Understanding the influence of socio-economic factors on emergency calls in The Hague

Group 22: Nicolò Canal (5625580), Rhys Evans (5633273), Matvei Isaenko (5618053), Thomas Sandbergen (4720814)
MSc Engineering and Policy Analysis
Technische Universiteit Delft, The Netherlands

***Abstract*—Since John Snow mapped cholera outbreaks in 1855 [1], spatial analysis has been widely used to study epidemiological phenomena and the general health of a population. Various research projects have investigated relationships between parameters like age [2], sex [3], and climate conditions [4] on the quality of provided medical services and, as a consequence, on the likelihood of serious health problems up to death [3]. This research's higher goal is to help the municipality of The Hague to provide equal access to emergency medical services despite the possible existing social and economic deprivation. To address this problem the spatial relationship between different types of socio-economic indicators throughout neighbourhoods in The Hague and the number of calls requesting emergency medical care in 2017 and 2018 was studied.**

***Index Terms*—data fusion, linear regression, ordinary least squares, k nearest neighbours, random forest, xgboost, geospatial**

I. Introduction

The Netherlands is often thought of as one of the wealthiest and most democratic countries in the world. According to the World Bank, the Netherlands is ranked 15th in GDP per capita [5] and it is ranked 9th on the Democracy Index [6]. Despite these positive rankings, the Netherlands has the second highest wealth inequality in the world after the United States [7]. This inequality is a complex phenomenon which influences a number of factors and population groups. This paper studies the influence of social and economic deprivation on ambulance calls in the city of The Hague. A number of different socio-economic indicators were taken into account to reflect this deprivation and to show geospatial distribution through neighborhood (buurt) level choropleths in The Hague. If resources are scare, then the most at-risk residents will be impacted disproportionately. The potential relationship between these indicators and geographical location of ambulance calls will give policymakers the insight to distribute resources equitably.

A. Research Question and Hypothesis

Based on this, our research question is: Are socio-economic indicators able to predict the distribution of ambulance calls among neighborhoods in The Hague? We will start our research by hypothesizing that socio-economic indicators alone will not be good predictors for the number of ambulance calls as they can, from our perspective, only offer a static, yearly overview of the situation in each neighbourhood while ambulance calls may require a more dynamic model with more granular data for indicators.

This report was prepared in partial fulfilment of the requirements for EPA1316 *Introduction to Urban Data Science* at Technische Universiteit Delft. Professor: Dr. Trivik Verma; TA Supervisor: Mobeen Nawaz.

II. Related Work

There has been previous research in the field of ambulance prediction. Butson et al. (1999) has been looking into the allocation of resources for ambulances [8]. More specifically it has been looking into estimating the cost of reducing response times for ambulances by adding more ambulances. This is done by developing a mathematical model to predict these response times. Furthermore, the paper also gives some more clear advice for allocating the resources based on the time-of-day and day-of-week and performing triage during the emergency calls. However, the paper does not look into whether it might be more efficient to allocate more resources to vulnerable groups.

Burt et al. have looked into this more specific ambulance data in the United States [9]. They have found that people aging above 75 are responsible of 26,3% percent of all ambulance trips. This number rises to 40,9% if we account for the ambulance trips that also brought the patient to the hospital. What we can also see is that women had just a little bit more ambulance trips than men, however they were hospitalized the same number of times. Furthermore, they seem to be more ambulance trips in the south of the US and in cosmopolitan areas. Last but not least there seem to be differences on ambulance trips in race and insurance status. Unfortunately, the papers states no reason for these differences.

A few reasons why some people are more vulnerable can be determined by race/ethnicity or education level. This effect can even be seen for multiple indicators for instance cervical cancer deaths or drug prescription. This might be explained that vulnerable groups in education level, race or income level may refrain from preventive measures which leads to higher costs [10].

However, because ambulance resources are allocated between communities, spatial information is also needed. Chandola (2012) has found that there is a clear link between spatial factors and health [11]. Spatial segregation between groups with different levels of wealth has an effect on health factors. Neighbourhoods with low income levels that are surrounded by other deprived regions have high mortality rates. However,

this can be mitigated in some areas where social support is abundant.

By using all this literature we can clearly see that there are some factors that are known to have an influence on our healthcare system i.e. age or education level. However there is a literature gap on the spatial variables that can help allocating ambulance resources. We simply want to understand if it is possible to use socio-economic variables can help us allocate these resources to specific neighbourhoods. Answering our research question will allow policymakers to mark certain areas which should receive more ambulance resources based on indicators with high feature importance. These areas should have certain characteristics in the found variables which indicate higher spatial vulnerability.

III. Exploratory Data Analysis

As already stated, this research will investigate whether we can use socio-economic indicators for predicting a spatial distribution of ambulance calls among neighborhoods of The Hague. During this research two different datasets will be used: the first one with data on ambulance calls and the second with the socio-economic data on neighborhoods (buurten) in The Hague. The first analysis is based on the ambulance calls in The Hague. The ambulance calls are plotted as points on the map of The Hague to give a general indication of the calls. The second step will be to merge both the ambulance and socio-economic data. Choropleths will give a spatial indication on whether the socio-economic variables can predict the ambulance calls. Last but not least a regression analysis will be performed. A model with socio-economic predictors and ambulance calls as a response variable will be trained and tested on the data of 2017. Additionally the model will be verified by predicting the 2018 data.

A. Data Description

As already said we are using two different datasets: one with data on ambulance calls in The Hague and the other one with socio-economic data. The first dataset has been provided to us by the lecturer. This dataset contains ambulance, police and firefighter calls collected in The Netherlands from January 2017 untill September 2020.

Data on socio-economic indicators were collected from the website Den Haag in Cijfers [12]. Den Haag in Cijfers offers abundant data on a smaller scale (buurten) in a relatively clean dataset. In this paper, socio-economic indicators were used to assess the inequality and deprivation that exist in society and to which vulnerable groups of the population are exposed. The process of selecting which variables to use started with an assumption that we want to have a list of variables that will reflect all major types of deprivation while keeping them useful for training of a regression model. The result of the selection is illustrated in Table I.

We did not use data for 2020, as they were seriously affected by the pandemic and related restrictive measures. We also limited our scope with taking into account only ambulance calls coming from the city of The Hague, because other services and geographical regions are not relevant for our research question. Data about calls made in 2019 remained unused as well. Even though these data are quite complete, several indicators from [12] were not available for 2019 at the moment this analysis was carried out.

Socio-economic Indicators			
№	Indicator Name	Description	Units
1	Number of ambulance calls	The total number of Priority 1 and Priority 2 ambulance calls in The Hague	# Calls
2	Gross population density	The number of residents per neighbourhood land area	# density
3	Average age of population	The mean age of residents in each neighbourhood	# years
4	Grey pressure	65+ as % of 15 to 64-year-olds	% people
5	Green pressure	-14 as % of 15 to 64-year-olds	% people
6	% Dutch natives	The percentage of people in a neighbour that are Dutch natives	% people
7	Low income private households	The percentage of households in a neighbourhood that have a lower than average household income	% households
8	Average income private households	The percentage of households in a neighbourhood that have an average household income	% households
9	High income private households	The percentage of households in a neighbourhood that have a higher than average household income	% households
10	Number of transportation and storage establishments	The number of establishments in a neighbourhood that belong to the transportation sector or storage sector	# Establishments
11	Number of medical facilities	The total number of medical facilities in a neighbourhood	# facilities
12	Gross non-housing density	The number of non-residential properties per neighbourhood land area	# density
13	Non-housing density (rateable value)	The number of non-residential properties per neighbourhood land area (rateable value)	# density
14	Gross housing density	The number of residential properties per neighbourhood land area	# density
15	Criminal offences	Number of criminal offences	# Crimes
16	% companies	Percentage of non-residential buildings that are companies	% properties
17	Number of people employed in construction industry	Number of people working in Construction	# People
18	Number of residents	The number of residents	# People
19	Buurt Index	Neighbourhood Index	n/a

Table I: List of socio-economic indicators

B. Data Cleaning

We can describe the EDA process used to create a 'clean' ambulance calls dataframe through the following steps:

1) Subset the dataframe to have only data about ambulance calls. Data is also subsetted to only include data of the municipality of The Hague.
2) The column named 'pmeTimeStamp' contained information about the year, month, day and time of a specific call. Knowing that each column should contain information about a single variable, we converted this column into four separated columns.
3) Choose only those columns that present useful information about ambulance calls for our analysis. We finally considered six variables: (1) Year; (2) Month; (3) Day; (4) Time; (5) Longitude; (6) Latitude.
4) Check for missing values, non-numerical coordinates, duplicated rows etc. 51 Rows had to be dropped because they were either duplicates or missed Longitude or Latitude data.

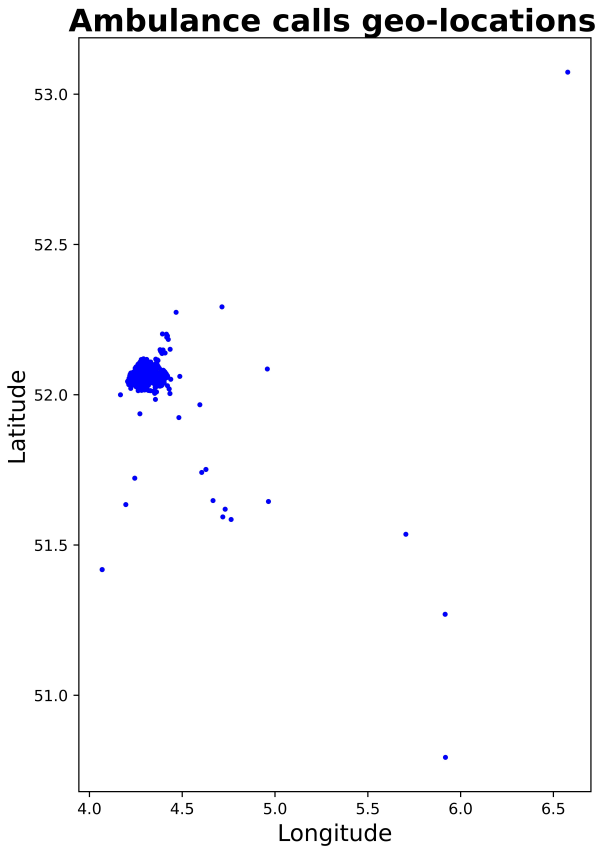

Figure 1. Initial geographical distribution of ambulance calls: each point corresponds to a location a call was made from. Outliers on this map indicate

5) Subset the data for the two years of interest for our analysis: 2017 for training our regression model; 2018 to evaluate the results of our model.
6) In the end of this first EDA, we obtained two dataframes: (1) Calls_2017; (2) Calls_2018. Each one of them contained five variables: (1) Month; (2) Day; (3) Time; (6) Longitude; (7) Latitude.
7) From this point the data cleaning is focused on the 'training year' 2017.
8) Some individual calls that are geographically not in The Hague are dropped (Fig 1). This was done by using the 'sjoin' function on the ambulance dataframe and the The Hague shapefile dataframe. This was verified by a geoplot (Fig 2)

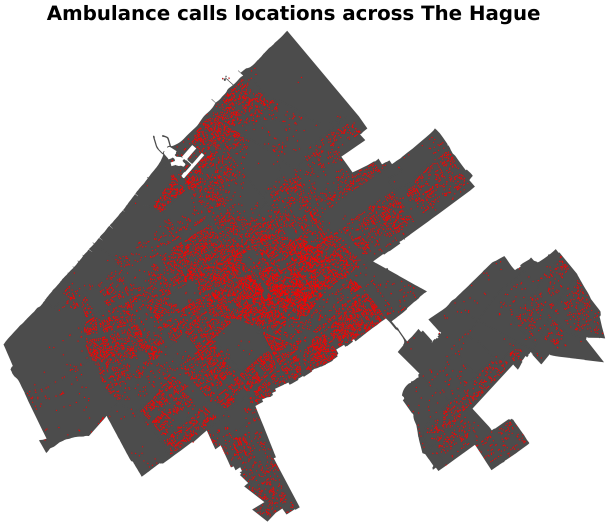

Figure 2. Geographical distribution of ambulance calls after excluding points outside of The Hague

9) Considering that our research goal concerns the spatial distribution of ambulance calls throughout the different neighbourhoods of The Hague, we created a first dataframe containing only two variables: (1) Neighbourhood; (2) Number of ambulance calls. The distribution of calls can be seen on the histogram (Fig. 3) and on the choropleth map (Fig. 4). In particular, we were able to identify three points in those neighbourhoods that majorly contributed to their unexpected number of calls (Fig. 5).

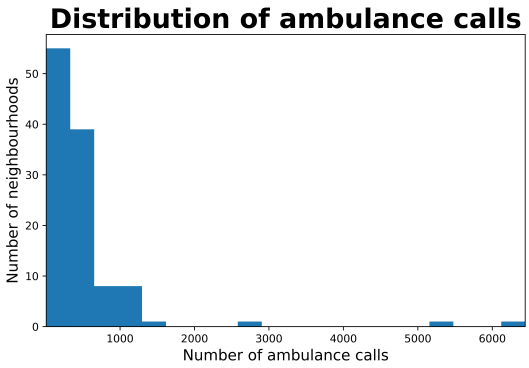

Figure 3. The initial histogram shows three anomalies

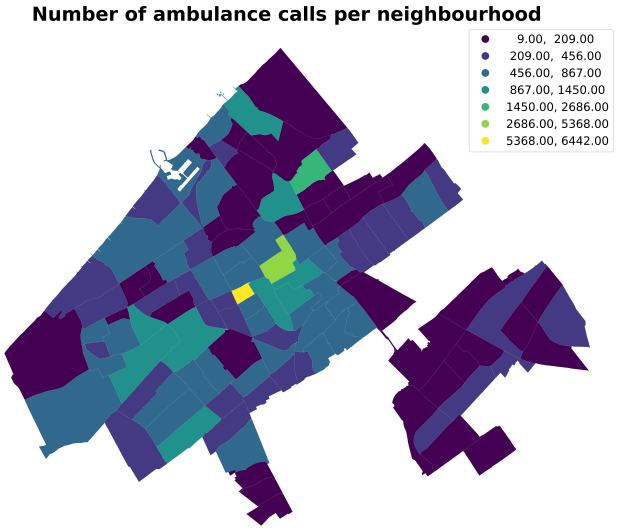

Figure 4. The initial choropleth map shows three neighborhoods that stand out among the others

10) Using the coordinates of those points, we used Google Maps to find out if they showed any kind of peculiarity. From that we understood that two of those points corresponded to two different hospitals in The Hague (HMC Bronovo and HMC Westeinde). The last point (52.069858, 4.291111) did not match any particular place (Fig. 5).
11) These 3 points are considered outliers and removed so they will not significantly affect the regression analysis. By doing that, the distribution of calls was greatly modified as we can see from Figures 6 and 7.

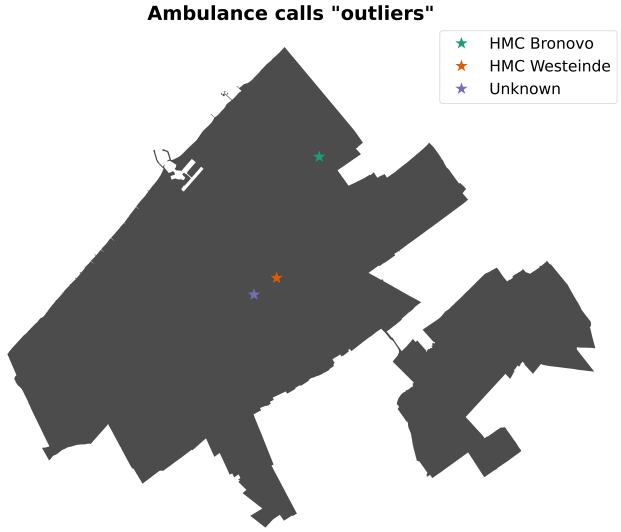

Figure 5. Founded sources of high number of calls from these points: 2 hospitals and one unknown location

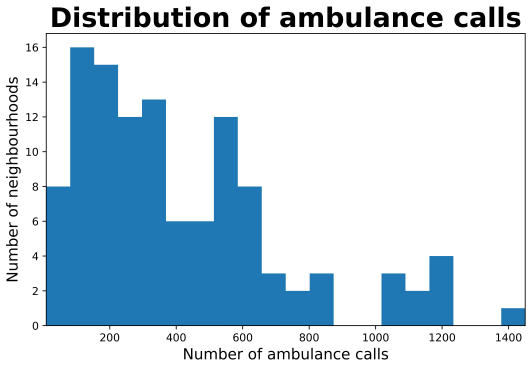

Figure 6. The histogram showed in figure 3 without anomalies

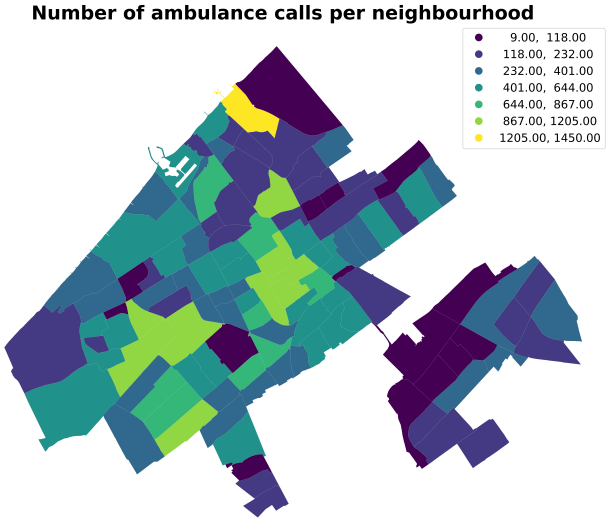

Figure 7. The choropleth map showed in figure 4 without anomalies

We can describe the EDA process used to create a 'clean' socio-economic dataframe through the following steps:

1) Data related to 2017 and 2018 were downloaded from Den Haag Cijfers [12].
2) Rows corresponding to Oostduinen, Vliegeniersbuurt and Tedingerbroek were removed. These neighborhoods have a very small or no population at all and therefore lack important data.
3) De Reef and Vlietzoom-West lack data on the neighborhood (buurt) level. The corresponding district (wijk) data from Hoornwijk is being used. The data that is used is an interpolation from the data for the years 2015, 2017 and 2018.
4) All columns end with —2017 or —2018. This suffix is removed.
5) The data is saved in a file for 2017 and a file for 2018.

C. Data limitations

The data we used has a few limitations:

1) The socio-economic data relies on population density and the amount of inhabitants. However in the real-world

there is a significant difference between suburban and business areas. Business areas have a low population density but a high amount of travellers from other parts of the city. These travellers will cause ambulance calls near their work location instead of near their house. Therefore there will be an imbalance in ambulance calls between suburban and business areas. Future research might solve this issue by assessing a daytime population of each neighborhood. For example, in [13] authors have calculated daytime population as the sum of the workplace population (employed people of age group 16-74) with residents younger than 16 and older than 74 year old. This aims to provide a proxy to the number of people active at a geographic area during working hours.

2) We are only taking the years 2017 and 2018 into account. The regression model might find weights that only apply to this limited dataset. Future research might expend on this by observing the situation over multiple years.
3) Our model was trained on a dataset that excluded points with anomalies described in section III-B, thus, we can not expect to correctly predict/evaluate the number of calls in the neighbourhoods that contain those points.
4) Although we have quite a lot of socio-economic variables, there are still some interesting variables. For instance the percentage of men and women or the average education level could be used. However if all relevant variables would be used, problems of multi-collinearity would appear. Future research might do more analysis using this variables and the best predictors from this research.

IV. Method

After merging the geographic shapefiles for The Hague with the selected socio-economic indicators, further analysis using Machine Learning becomes possible. To conduct this analysis, we chose to implement and compare several regression models. With this regression model we can evaluate whether the socio-economic indicators are able to predict the distribution of ambulance calls, which is the goal of this research.

The following supporting points are in favor of using regression models:

1) First of all a socio-economic variable with a small regression coefficient will not have much influence on the response variable. This will mean that there are other more important socio-economic variables. However we have to take the size of the units of the variable into account.
2) Furthermore we can evaluate the regression model by using evaluation methods. If for instance, the Mean Squared Error is quite high, the socio-economic variables will not be able to predict the ambulance calls very well.
3) Last but not least the regression model can take spatiality into account. This will tell us if we can also predict the distribution of ambulance calls on a spatial level.

A. Regression Models

For this report, seven models are implemented using four different regression methods. To start, a model is developed using Ordinary Least Squares (section V-B). The indicators used in this model are chosen by looking at the correlation between the indicators and the number of ambulance calls in each neighbourhood (section V-A). Furthermore a k Nearest Neighbor regression model has been performed (section V-C) to take spatiality into account.

After developing both of those models, we decided to explore further regression techniques to see how they compared with our two previous approaches. We develop a model using Random Forest regression (section V-D) and a model using XG Boost regression (section V-E). Random Forest regression is an ensemble learning method that works using decision trees. Since it is an ensemble learning method, it uses multiple algorithms to hopefully obtain better results. XGBoost is overkill for this situation, but we implemented it to see how it compares to more simple regression methods. XGBoost is a very popular method and is generally considered a ”state-of-the-art” machine learning algorithm.

B. Shapley Values for Feature Importance

To determine which indicators are most impactful, Shapley values are looked at. Shapley values are one of the most commonly used methods for determining the importance of an indicator in a regression model. They are very helpful in promoting explainable AI as they can clearly show the effect an indicator has on the model [14]. We will use Shapley values in our analysis to see which socio-economic indicators impact our regression models the most. This information is relevant as it could be used to help guide policy based on which indicators have the highest feature value.

V. Analysis

This chapter will discuss all relevant findings from the analysis. These findings will be explained by relevant figures from for instance the regression models.

A. Correlation between variables

Based on our analysis we have found that there is some correlation between our variables. This can be seen in Figure 8. The variables with the most correlation are ’Average age of population’ and ’Grey pressure (65+ as % of 15 to 64-year-olds’). However these variables won’t cause the problem of multi-collinearity as they indicate different characteristics of the neighbourhood. These variables explain the same phenomenon as by their definition. Therefore they won’t affect each other and thus not the predicitve value of the model.

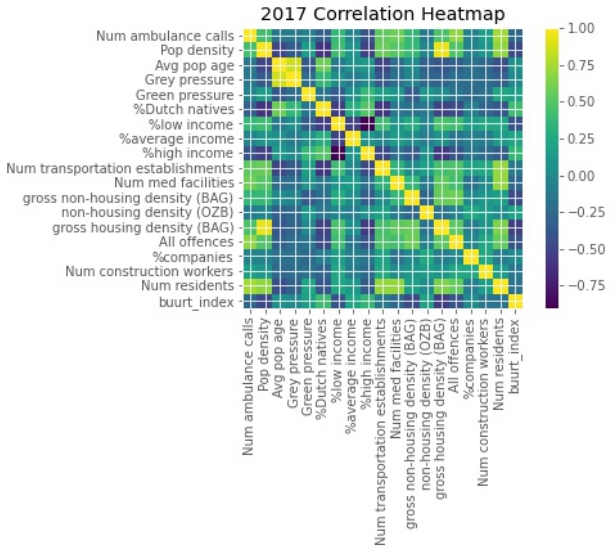

Figure 8. Correlation Heatmap 2017 [15]

B. Ordinary Least Squares Regression

A model is developed using Ordinary Least Squares. The indicators used in this model are chosen by looking at the correlation between the indicators and the number of ambulance calls in each neighbourhood. Figure 8 shows the correlation heatmap for 2017. From this heatmap, we were able to rank the indicators by their correlation to the number of ambulance calls. This analysis showed us that the four most correlated indicators were

1) Number of residents in each neighbourhood
2) Number of criminal offences in each neighbourhood
3) Number of medical facilities in each neighbourhood
4) Number of transportation and storage establishments in each neighbourhood

OLS Regression Formula

Number of ambulance calls $\sim$

Number of residents +

Number of criminal offences +

Number of medical facilities +

Number of establishments, transportation and storage

This OLS regression method resulted in an R^2 value of 0.612

C. k Nearest Neighbour Regression

Next, a model using k Nearest Neighbour regression was developed. Unlike the Ordinary Least Square regression model, this model considered all of the aforementioned socio-economic indicators. In order to determine the best number of neighbours to use for this model, a simple analysis was done in Figure 9 to find the k that corresponded with the highest R^2.

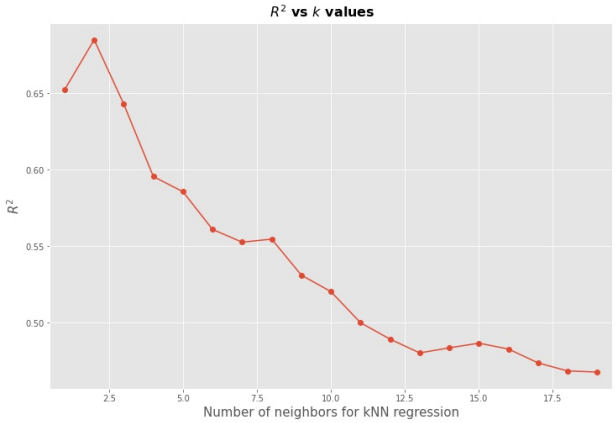

Figure 9. R^2 value as k increases in k NN [16]

Figure 9 shows a sharp increase in the R^2 value as the k increases initially. It reaches its peak—and presumably absolute maximum–when $k = 3$ and then decreases from there. Therefore, $k = 3$ and three neighbours are used for the k Nearest Neighbour regression model.

The k Nearest Neighbour regression model resulted in an R^2 value of 0.713

D. Random Forest Regression

After successfully creating a model with a high R^2 score, we were interested in seeing what other types of regression algorithms were capable of. Therefore, we created two models using Random Forest Regression. The first of these methods once again considers all of our selected indicators.

This first Random Forest regression model resulted in an R^2 value of 0.504. Since this R^2 value was lower than the R^2 obtained using k Nearest Neighbour regression, we were interested in finding out which indicators impacted this new Random Forest regression model the most. To do this, we employed Shapley values to discover the feature importance.

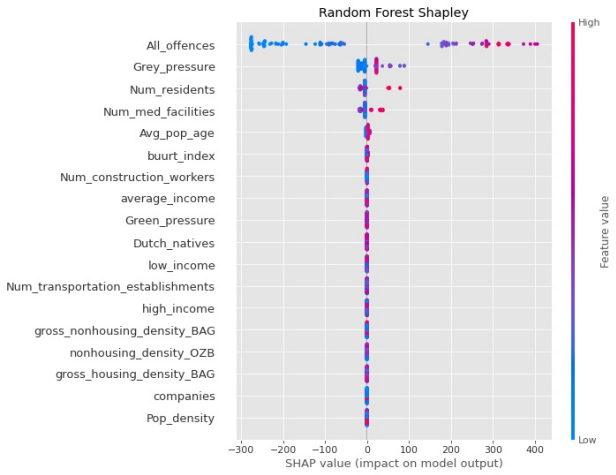

Figure 10. Shapley values for feature importance for Random Forest regression [17] [16]

Figure 10 illustrates the feature importance for the Random Forest regression model. This analysis clearly displays the overpowering influence the number of criminal offences in a neighbourhood has on the model. This is interesting to us as researchers as the link between the number of criminal offences and the number of ambulance calls may not immediately be clear. Presumably, most ambulance calls are not to treat victims of heinous crime as The Hague is a relatively safe city. Therefore, we as researchers must try to draw other connections and identify the causal links that results in neighbourhoods with more criminal offences also having more ambulance calls.

To further explore this case, a second model using Random Forest regression was created except this time the number of criminal offences was not taken into consideration as an indicator.

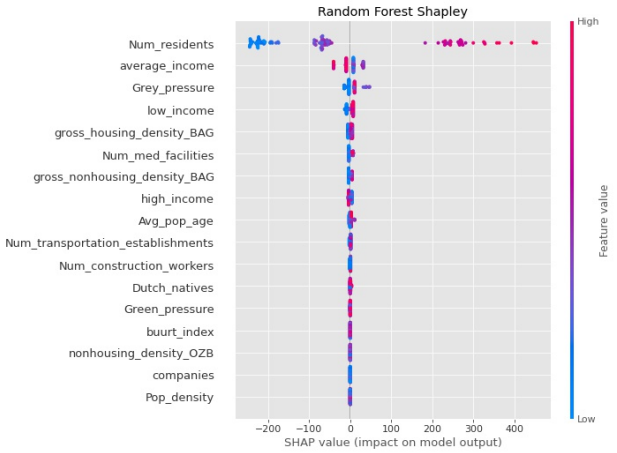

Figure 11. Shapley values for feature importance for Random Forest regression without crime as an indicator [17] [16]

This model, as show in Figure 11, was heavily impacted by the number of residents in each neighbourhood, absent the number of criminal offences indicator. This correlation may be conceptually more intuitive as it makes logical sense that neighbourhoods with more residents have more ambulance calls. However, this may not always be the case as the most visited neighbourhoods in a city will likely have more emergency calls even though they have less permanent residents.

This second Random Forest regression model (without criminal offences) resulted in an R^2 value of 0.525

E. XG Boost Regression

The last regression method we chose to explore is XG Boost regression. While this approach may be overkill for our situation, it is considered a state-of-the-art regression algorithm and we were interested in seeing how it compared to our previous models.

Surprisingly, this first XG Boost regression model yielded the lowest R^2 value yet at 0.292

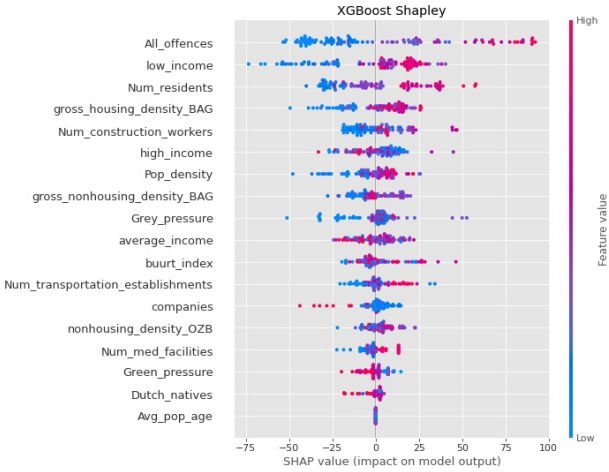

Figure 12. Shapely values for feature importance for XG Boost regression [17]

Again, we looked at the Shapley values for feature importance to see what impact each indicator had compared to the Random Forest models. In Figure 12, we see that the impact of one particular indicator is less clear than in the Random Forest models. Despite this, the number of criminal offenses still reigned supreme in terms of feature value. It again appeared to have the highest impact on the model. Because of this, we again decided to create another model using the same algorithm, but this time without the number of criminal offences being considered as an indicator.

Figure 13. Shapely values for feature importance for XG Boost regression without crime as an indicator [17]

Figure 13 shows the Shapley values for feature importance of the XG Boost model without the number of criminal offences being considered as an indicator. This time, we see that the percentage of households in a neighbourhood that have a low income was the feature of highest value. This could potentially be due to a link between low-income and crime, but more research will need to be done to see if that link holds true in The Hauge.

The R^2 value of this second XG Boost model, this time without criminal offences, was 0.325.

F. Linear Regression Based on Shapley Feature Importance

Since the number of criminal offences was the feature with the most importance in the Random Forest and XG Boost regression models, we decided to create a simple linear regression model using the number of criminal offences alone (as the only indicator) to predict the number of calls in a neighbourhood. The plot generated from this analysis is shown in Figure 14.

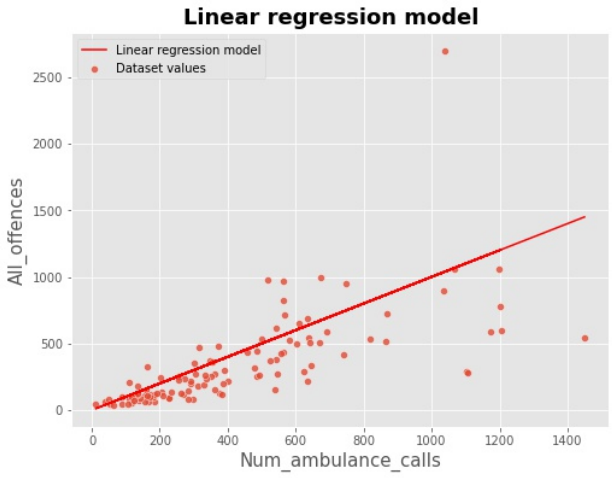

Figure 14. Linear regression model predicting the number of ambulance calls using the number of criminal offences alone

Figure 14 shows the linear correlation between the number of ambulance calls in a neighbourhood and the total number of criminal offences in that neighbourhood. Some notable outliers are visible, however.

This linear regression model, using only the number of criminal offences, generated the best R^2 value with a score of 0.759.

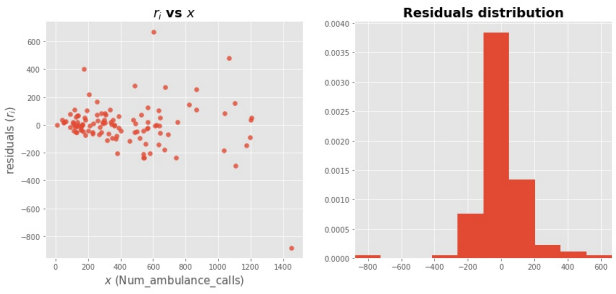

Figure 15. Residual plot for the linear regression model shown in Figure 14

The residual plot shown in Figure 15 appears Normal for the linear model using only the total number of criminal offences.

VI. Discussion and Conclusion

From the regression models constructed during this report, it is evident that a link exists between socio-economic factors and the number of ambulance calls in neighbourhoods in The Hague. Surprisingly, the simple linear regression model that only looked at the number of criminal offences in a neighbourhood resulted in the best R^2 value with a score of 0.759. However, this high R^2 value could potentially be a result of over-fitting. Future research will need to be done when ambulance call data from more years is available for training and testing to see if the model's high predictive score holds up. A major limitation of this analysis is that data from only one year is used for training and data from only one year is used for testing. Ideally, we would train on data from several years and then have several other years to use to evaluate the model's predictions. Despite this, the models created in this report offer potential insight into the impact different socio-economic indicators have on the number of ambulance calls in neighbourhoods throughout The Hague and provides insightful analysis to help guide public policy in the future.

A. Strongest predictors

It can be seen in the Random Forest Regression in Figure 10 that the variable 'All offences' is the strongest predictor. This means that a small change in 'All offences' will directly result in a significant change in ambulance calls. This trend continues in XGBoost in Figure 12 although less prominent.

If we remove the variable 'All offences' as in Figure 11, we can see that the number of residents living in the area is the strongest predictor. This information is interesting because one might think that population density would have a higher impact on the model. We previously hypothesized that we would have to take the difference between suburban and high-density areas into account (section III-C) but the Shapley values for the Random Forest regression models appear to disagree with this. Figure 11 shows that the total number of residents in a neighbourhood is the feature of highest value while the population density of a neighbourhood is the feature with lowest value.

B. Future research

While our research is a good first step, future research could expand on our approach to offer more insight for policymakers. Taking into account all the different limitations of our presented research models, we have formulated three main future research possibilities:

1) Planning ambulance resources throughout The Hague should not only consider the spatial distribution of ambulance calls across the city. If the final goal of the municipality is to optimise the allocation of ambulance resources in the city, time distribution of the calls should also be taken into consideration. Identifying 'peak periods' of ambulance calls throughout the year is crucial to understand if the allocated capacity of ambulances will be able to cope with these peaks. Nevertheless, to achieve this last objective we should introduce new indicators/parameters, since the socio-economic indicators used throughout our analysis only capture 'yearly' pieces of information about the city.

Thus, if we want to describe the 'time distribution' of ambulance calls throughout the year we should consider indicators that have a more granular time definition. To reach this last goal we could, for instance, gather information on weather, traffic, public transports usage and social media data from the city.

2) In our research, we could not analyse the efficiency of ambulance services across The Hague. In other words, due to the lack of data in the provided ambulance calls dataframe, we could not understand how much time is required for an ambulance to reach different points in The Hague. This last analysis would be of great interest to understand if there are some 'blind spots' in the city that are associated to greater time periods for an ambulance to reach them. Finding these 'blind spots' would be crucial, especially if they fall inside neighbourhoods with large amounts of ambulance calls, to potentially organise an optimised distribution of ambulance resources across the city.
3) Currently our socio-economic indicators are not capable of describing the dynamics of every day life in the city. For example, several areas of The Hague have a lower permanent population, but have a high daily population due to workers and tourists. This influx is dependent on time and could be better modeled using time-series data at a more granual hour-by-hour level throughout the day. Therefore, while our models are correct, they are only able to capture a static image of the city. This can be improved by creating a dynamic model using a time series analyis. Exploratory data analysis for this approach has been explored in research by Srivastrava et al. and it may be promising for future research [18].

REFERENCES

[1] J. Snow, *On the mode of communication of cholera, by John Snow ... 2d edition, much enlarged*. J. Churchill, 1855.

[2] L. K. S. Sylvia Bernard, "Emergency admissions of older people to hospital: a link with material deprivation," *Journal of Public Health*, vol. 20, pp. 97–101.

[3] T. Smith, "The relative effects of sex and deprivation on the risk of early death," *Journal of Public Health*, vol. 14, p. 402–407.

[4] S. C. Timothy J.Dolney, "The relationship between extreme heat and ambulance response calls for the city of toronto, ontario, canada," *Environmental Research*, vol. 101, pp. 94–103.

[5] "Gdp per capita, ppp (current international $) — data," https://data.worldbank.org/indicator/NY.GDP.PCAP.PP.CD?year_high_desc=true, (Accessed on 11/04/2021).

[6] "Democracy index 2020 - economist intelligence unit," https://www.eiu.com/n/campaigns/democracy-index-2020/, (Accessed on 11/04/2021).

[7] C. Balestra and R. Tonkin, "Inequalities in household wealth across oecd countries: Evidence from the oecd wealth distribution database," *OECD Statistics Working Papers*.

[8] A. J. Fischer, P. O'Halloran, P. Littlejohns, A. Kennedy, and G. Butson, "Ambulance economics," *J Public Health Med*, vol. 22, no. 3, pp. 413–421, Sep 2000.

[9] C. Burt, L. Mccaig, and R. Valverde, "Analysis of ambulance transports and diversions among us emergency departments," *Annals of emergency medicine*, vol. 47, pp. 317–26.

[10] A. Perzynski, S. Shick, and I. Adebambo, *Health Disparities Weaving a New Understanding Through Case Narratives: Weaving a New Understanding Through Case Narratives*, 01 2019.

[11] T. Chandola, "Spatial and social determinants of urban health in low-, middle- and high-income countries," *Public Health*, vol. 126, no. 3, pp. 259–261, Mar 2012.

[12] "Den haag in cijfers - databank," https://denhaag.incijfers.nl/jive, (Accessed on 11/02/2021).

[13] A. Noulas, C. Moffatt, D. Hristova, and B. Gonçalves, "Foursquare to the rescue: Predicting ambulance calls across geographies," *ACM Digital Health'18*.

[14] C. Harris, R. Pymar, and C. Rowat, "Joint shapley values: a measure of joint feature importance," 2021.

[15] T. pandas development team, "pandas-dev/pandas: Pandas," Tech. Rep., Feb. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3509134

[16] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[17] S. Lundberg, "An introduction to explainable ai with shapley values — shap latest documentation," https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html, 2018, (Accessed on 11/02/2021).

[18] V. Srivastrava, M. Jhatakia, M. Subramanian, and Z. Jamadar, "Predicting the number of ambulance calls in the hauge using socio-economic indicators aggregated at the neighbourhood level," 2021.

APPENDIX A
INDICATOR CHOROPLETHS

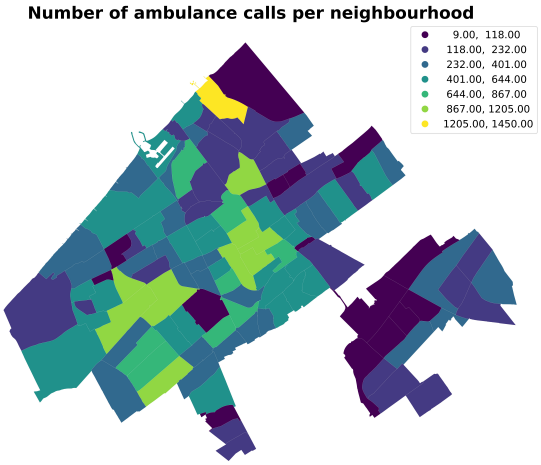

Figure 16. Number of ambulance calls per neighbourhood

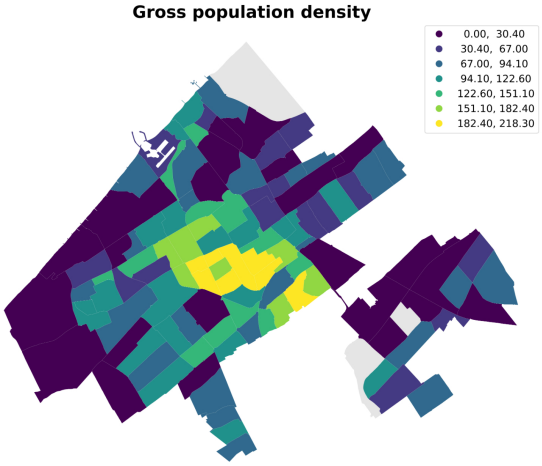

Figure 17. Gross population density per neighbourhood

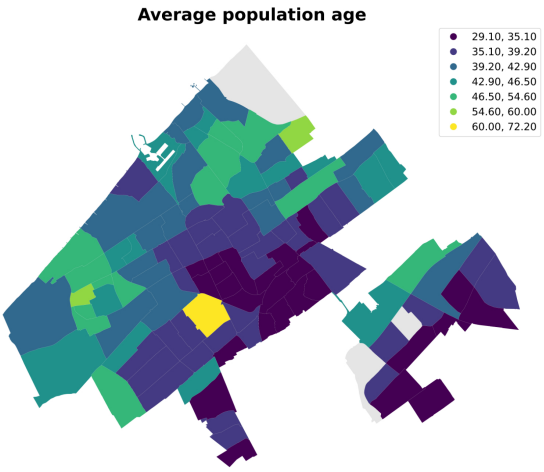

Figure 18. Average population age per neighbourhood

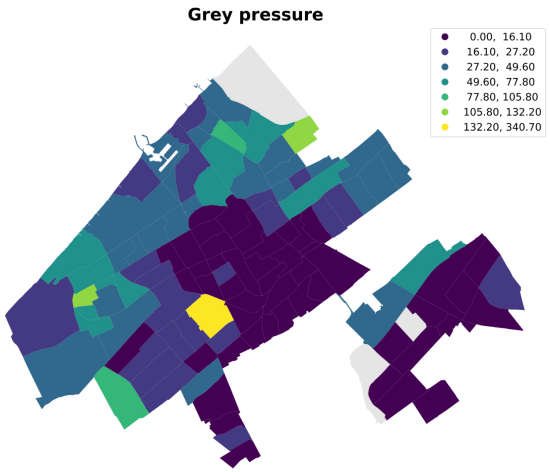

Figure 19. Grey pressure per neighbourhood (65+ as % of 15 to 64-year-olds)

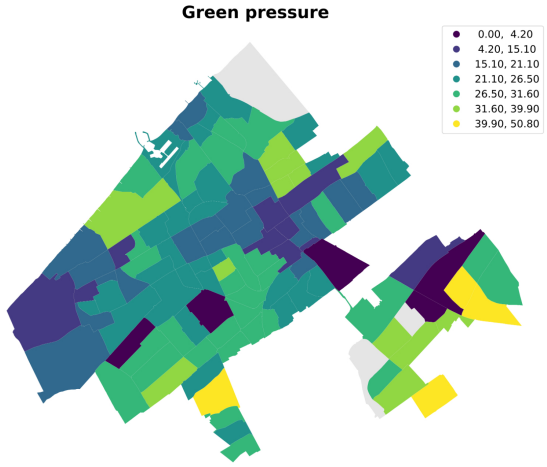

Figure 20. Green pressure per neighbourhood (-14 as % of 15 to 64-year-olds)

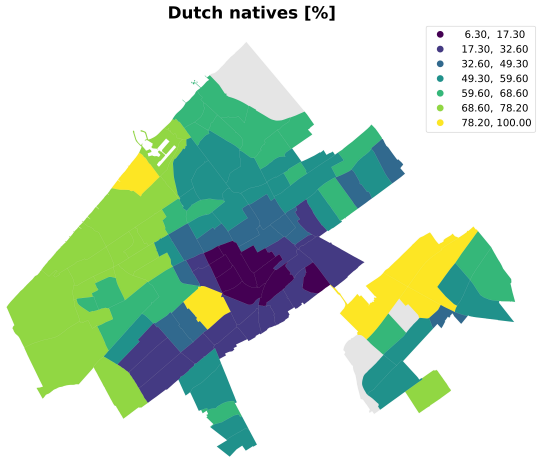

Figure 21. Percentage of population that is a Dutch native in each neighbourhood

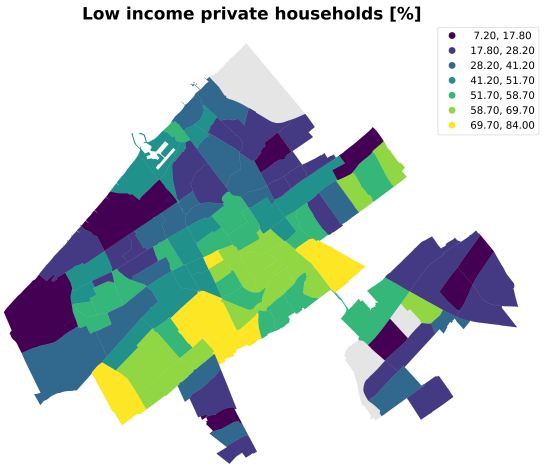

Figure 22. Percentage of households with low income in each neighbourhood

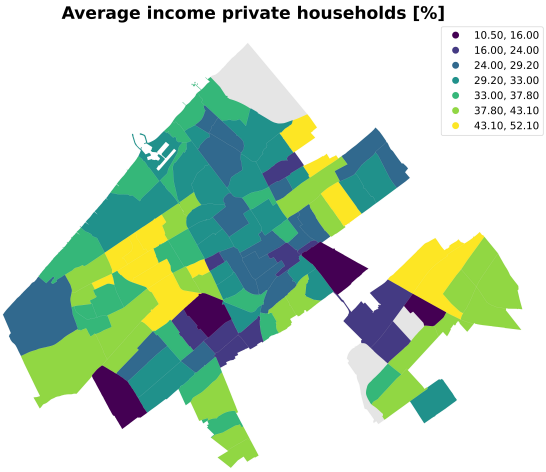

Figure 23. Percentage of households with average income in each neighbourhood

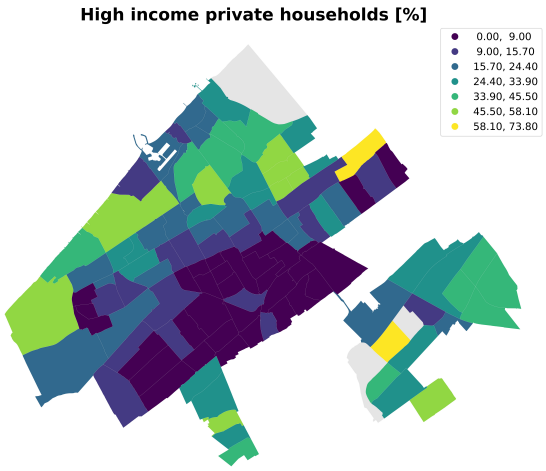

Figure 24. Percentage of households with high income in each neighbourhood

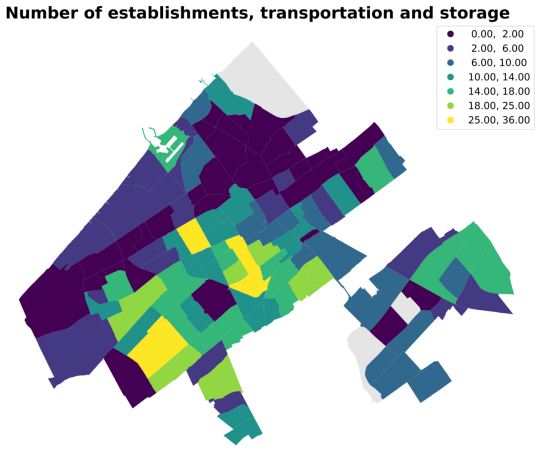

Figure 25. Number of transportation and storage establishments in each neighbourhood

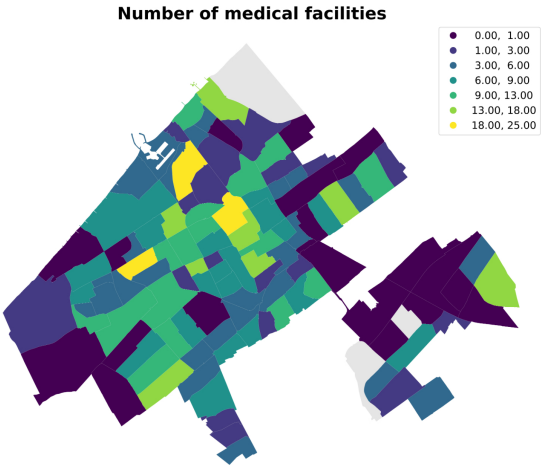

Figure 26. Number of medical facilities in each neighbourhood

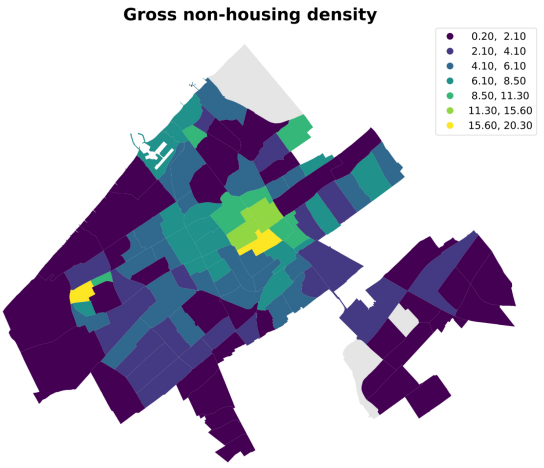

Figure 27. Gross density of non-residential properties per neighbourhood

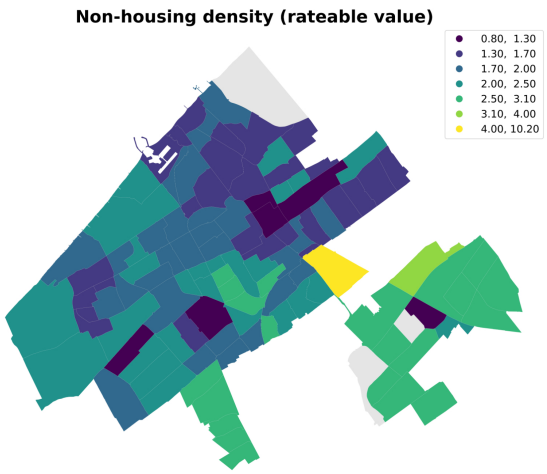

Figure 28. Average number of non-residential properties per hectare

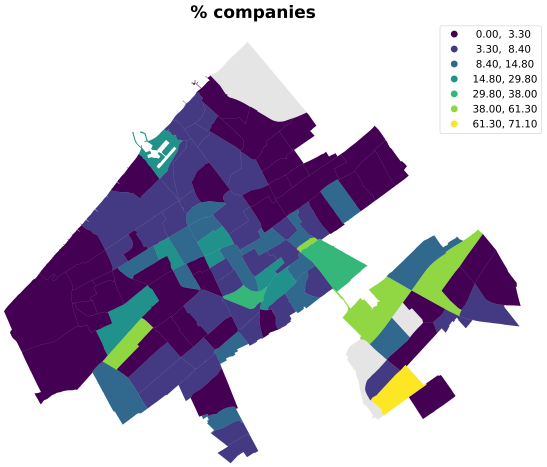

Figure 31. Percentage of non-residential buildings that are companies

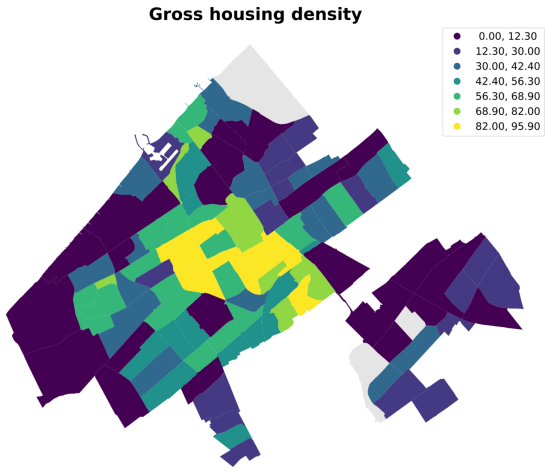

Figure 29. Gross density of residential properties per neighbourhood

Figure 32. Number of people employed in construction industry in each neighbourhood

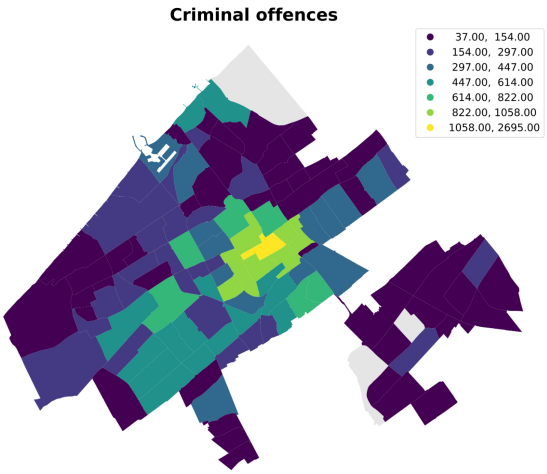

Figure 30. Number of criminal offences

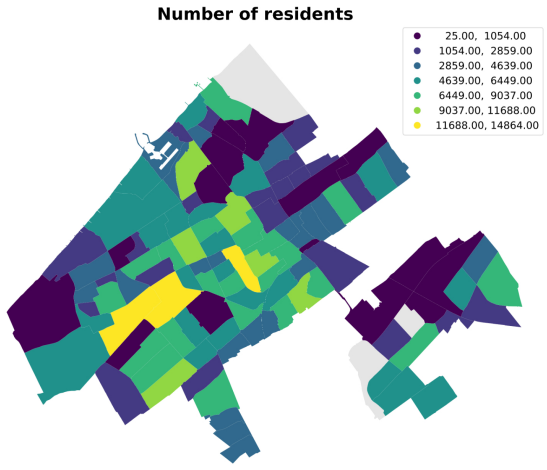

Figure 33. Number of residents in each neighbourhood