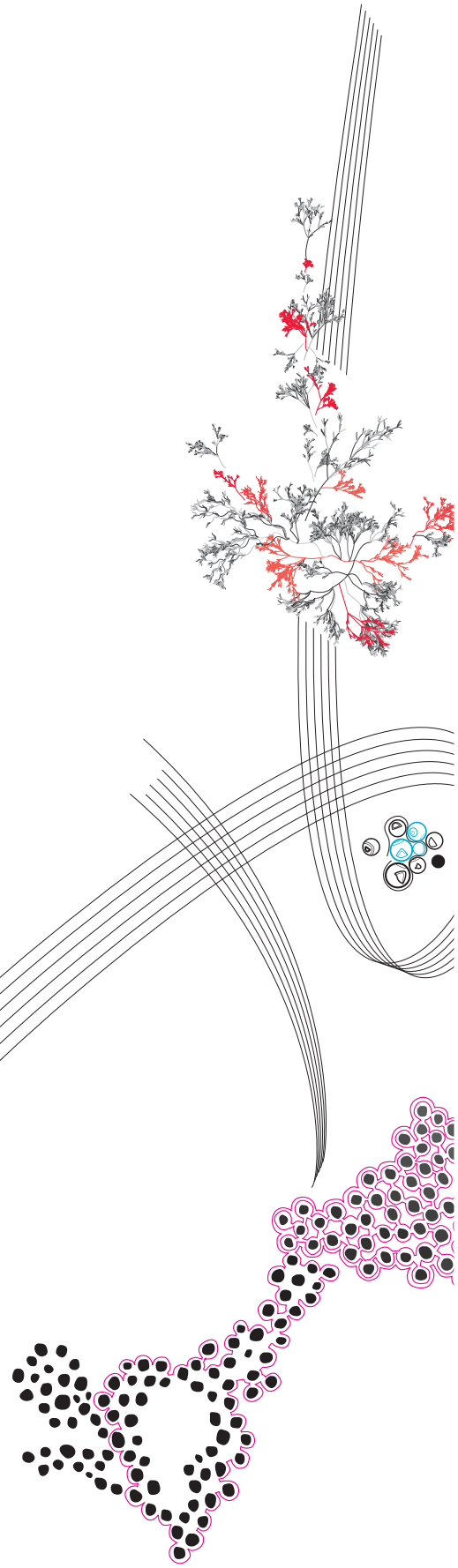

BSc Thesis Applied Mathematics

Spatio-Temporal Analysis on Crime Occurrences in Enschede and Campus Area

A. Wilbrink

Supervisor: prof. dr. M.N.M. van Lieshout

July 11, 2025

Department of Applied Mathematics
Faculty of Electrical Engineering,
Mathematics and Computer Science

UNIVERSITY OF TWENTE.

Preface

This thesis is one of the final steps in my Bachelor's programme in Applied Mathematics at University of Twente. The whole study always was challenging, but was always satisfying. Every module, there was space to apply mathematical tools to a real-world problem.

When we could choose for the themes of our thesis, the topic of this research took me immediately, among else. This was because I really like to apply the mathematics to real world problems. Also, I like the combination of statistical modelling, spatial data, and social context.Throughout the project, I learned very much, both about writing a research paper and about statistics again. Also about working with real data, and how you interpret results carefully.

I want to thank my supervisor, prof. dr. M.N.M. van Lieshout, for her guidance, encouragement, and many helpful suggestions throughout this thesis project. Her expertise and support every time helped me further when I was stuck in a big part of literature. Thanks for introducing me to the field of Generalized Linear Models.

I also hope this thesis reflects the knowledge and skills I have gained during my studies.

Enschede, July 2025

Bart Wilbrink

Spatio-Temporal Analysis on Crime Occurrences in Enschede and Campus Area

A. Wilbrink

July 11, 2025

Abstract

This thesis presents a spatio-temporal analysis of the assaults in the city of Enschede. The main goal is to find factors that contribute the number of assaults in Enschede. Data is taken from the Dutch police, and CBS (Centraal Bureau van de Statistiek).

In this research, the question *Is there a shift in occurrences of assaults in Enschede?* is addressed and split up in three sub research questions. The results of these research questions show that population, rental housing, low income households, and hospital-related businesses are good explanatory variables. We can unfortunately not take a big conclusion about the number of incidents in Enschede-Noord, the campus area.

This study shows how modelling can be used for public safety. In the conclusion, recommendations are made for further research. *Keywords*: Spatio-temporal modelling, Poisson regression, crime data, assaults, neighbourhood analysis, statistical modelling, Enschede, socio-demographic factors

1 Introduction

Violence in public space has a large impact on society. Besides the direct harm it causes to the victims, it also gives a general feeling of unsafety. Inhabitants do not feel safe in their own neighbourhood, which can also cause that cases reinforce each other. Therefore, it is important to understand where these incidents happen, and what the causes are. If policymakers know this, they can make better policy. Recently, on the campus, a series of these incidents have happened.[1] This caused some unrest in the statistics research group of the University of Twente.

Last 10-15 years, more and more public data is available about neighbourhoods, crime and cities. This makes it better possible to study crime patterns. In this report, we do this by taking into account the place and the time, which is *spatio-temporal modelling*. In this thesis, we focus on the municipality of Enschede between 2013 and 2024. Using data from CBS and the Dutch police, we want to find trends and want to know what influence violence in public space. We do this using Poisson regression, a technique which is very good useful method for modelling count data. In this research we also include population size, income level, level of rental houses and other socio-demographic factors. Also the neighbourhood of Enschede-Noord is studied, if there is unusual development over time. The term *assaults* will be used for violent incidents that occur in public space.

The main research questions of this research is: *Is there a shift in occurrences of assaults in Enschede?*. Because this is quite a broad question, we begin by making the model, so we want to answer the following question: (1) *What factors influence the number of assaults?* If we know this, we can apply this model, and look if there is a unusual development in

Enschede-Noord: (2) *For the neighbourhood of Enschede-Noord, is there a big growth in number of assaults last year?* Lastly, we also want to know something about Enschede in general, which we handle by answering the third research question: (3) *What can we say about the temporal trend of the assaults?*

By answering these questions, we hope to give a better understanding of violence in public space.

2 Data and data preparation

2.1 Occurrence data

To answer the research questions presented in Section 1, data is needed. The data used in this report are retrieved in CSV format from Politie Nederland [2], which provides crime statistics every month through its open data portal. This data set contains monthly records of various types of crimes, reported at both the neighbourhood and district level. Districts are smaller areas within neighbourhoods, so every neighbourhood is subdivided into different districts. The data is from January 2012 to March 2025 and is recorded on a monthly basis. This makes we can use this perfectly for spatial and temporal analysis. Monthly updates make sure that the dataset remains up to date. Also recent developments can be researched quick. Because the statistics are published by the police based on official crime reports, this dataset can be considered reliable for this analysis.

The dataset contains more data than we really need, so before proceeding with the analysis, we want to clean the dataset. We only consider assaults in public space, so it is necessary to select the only relevant types of crime. We filtered the dataset by the following five crime categories: (1). Openlijk geweld (public violence), (2). Bedreiging (threat) (3). Mishandeling (assault) (4). Straatroof (mugging) (5). Overval (robbery).

These five categories at least cover all the cases we want to address, as they typically happen in public space. By summing up over these categories, we can list the number of assaults for every month from January 2012 up to and including March 2025.

Before downloading, filtering by crime type and city is already possible. After loading the dataset in R, we can view what it looks like, see Table 1. In this dataset, "Totaal misdrijven" (total crimes) is also included, which is the total number of crimes that month in a given neighbourhood. For the analysis of the assaults in public places, we will not need this record. Across the analysis, we also came through some inconveniences in the dataset. We for example had to replace a "," by a "." and make the data numeric to fit correctly. Also, the year numbers had to be factorized.

Crime type	**Period**	**Area**	**Reported crimes**
1.4.3 Openlijk geweld (persoon)	2017 juli	Wijk 00 Binnensingelgebied	1
1.4.4 Bedreiging	2017 juli	Wijk 00 Binnensingelgebied	7
1.4.5 Mishandeling	2017 juli	Wijk 00 Binnensingelgebied	19
1.4.6 Straatroof	2017 juli	Wijk 00 Binnensingelgebied	0
1.4.7 Overval	2017 juli	Wijk 00 Binnensingelgebied	0
Totaal misdrijven	2017 juli	Wijk 00 Binnensingelgebied	350
1.4.3 Openlijk geweld (persoon)	2017 juli	Wijk 01 Hogeland - Velve	0
⋮	⋮	⋮	⋮

Table 1: Some rows from the filtered dataset from Politie Nederland

2.2 Geographical data

To visualise the data spatially, we use the boundary data provided by Centraal Bureau voor de Statistiek (CBS) [3]. These are part of the geopackage with data from 2023. It contains polygon shapefiles for neighbourhoods, districts and municipalities across the Netherlands. A part of this table can be seen in Table 2.

Buurtcode	**Buurtnaam**	$\cdots$	**Geom**
BU01530000	City	$\cdots$	MULTIPOLYGON(((...)))
BU01530001	Lasonder, Zeggelt	$\cdots$	MULTIPOLYGON(((...)))
BU01530002	De Laares	$\cdots$	MULTIPOLYGON(((...)))
$\vdots$	$\vdots$	$\ddots$	$\vdots$
BU01530908	Buurtschap Twekkelo	$\cdots$	MULTIPOLYGON(((...)))

Table 2: Some rows and columns from the geopackage from CBS

This boundary data is from 2023. Although boundary data is available for each year, for simplicity, we assume that the boundaries have remained unchanged for every year. This is a reasonable assumption, as we only use data from the last 12 years. As a result, we can use the geometry column in the 2023 geopackage as boundary data for the whole of our research.

The geopackage consists of several layers: *buurten* (districts), *wijken* (neighbourhoods) and *gemeenten*(municipalities). If we want to research or visualize on a neighbourhood scale, we use layer *wijken*. But if we want to look closer, we use layer *buurten*. We use this dataset for visualization of the crime data. We can match the area names in the crime dataset with those in the boundary dataset. With this, we can map occurrences to areas and make visualisations.

2.3 Socio-demographic data

Now that we have the crime data, we also want to find data which can be used as explanatory data. CBS also provides annual data about neighbourhoods and districts, called *Kerncijfers wijken en buurten* [4]. These datasets contain a wide range of variables for each neighbourhood and district in the Netherlands. We use the datasets from 2013 through 2024, as data from 2012 is formatted in another style, which makes it very difficult to fit to the new datasets. A part of this dataset is shown in Table 3. In total there are around 125 variables in this dataset. However, many of them are empty, so we can not use every variable. At the end of this section, we will explain how we deal with that.

Neighbourhood	**Number of inhabitants**	$\cdots$	**Rate of urbanization**
Wijk 00 Binnensingelgebied	28505	$\cdots$	1
Wijk 01 Hogeland - Velve	12505	$\cdots$	2
Wijk 02 Boswinkel - Stadsveld	21525	$\cdots$	1
Wijk 03 Twekkelerveld - T.H.T.	9035	$\cdots$	2
$\vdots$	$\vdots$	$\ddots$	$\vdots$

Table 3: Some rows and columns from the neighbourhood information dataset, 2013

We load each annual dataset in R, and polish it before doing analysis. Across years, the structure of the dataset still varied. However this only required small modifications.

For every dataset, we filtered on the type of the region, which should be `neighbourhood`, as we will do an analysis on neighbourhood level. Also, we added a new column, called `year`, to indicate which year this data is about. These yearly datasets were then combined into one dataset.

Due to changes in the structure of the datasets over the years, some variables appeared under different column names. For example, the number of rental homes was categorized under `wonen` in some years, and under `wonen_en_vastgoed` in others. After confirming that these variables were the same, we merged the two columns. Once we did this for all duplicate variables, we removed variables that contained missing values, to ensure consistency. Finally, we then have a clean set of explanatory variables. In Section 4.1, we will explain which covariates will be used.

2.4 Exploratory data analysis

Before building the model and selecting covariates, one has to take a close look at the data set. Doing this exploratory data analysis gives us insight in the structure, trends and relationships in the data. It also helps to detect missing values or outliers. By getting an initial understanding of the dataset, we can make better decisions about which variables to include and how to model them (for example, use log). Therefore, in this section, we will first look at temporal trends, followed by spatial trends or relationships between neighbourhoods. This forms a foundation for the development of the mathematical model in Section 3.

2.4.1 Temporal trends

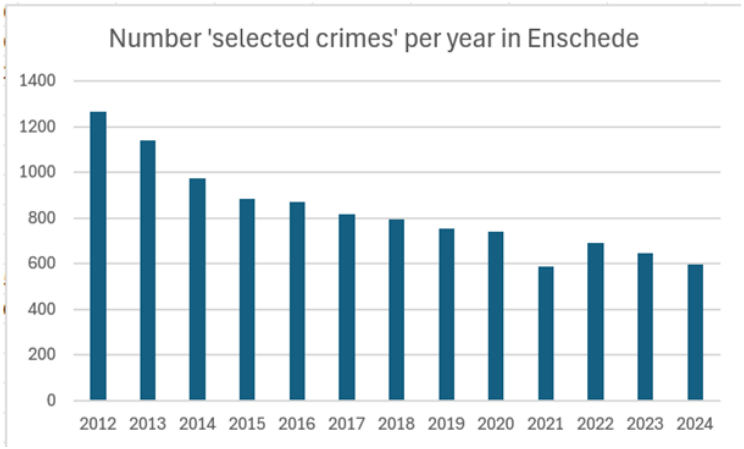

FIGURE 1: Number of assaults per year in Enschede

To understand how crime has changed over time in Enschede, we analyse both yearly and monthly patterns. Figure 1 shows that the annual number of assaults in Enschede has clearly decreased from 2012 to 2024. Zooming in on the monthly data in Figure 2 shows more short-term variability. For example, each January consistently shows a low point. We can think about potential underlying causes, to account for in the analysis. Possibly, a seasonal component may be useful when modelling monthly data. In Figure 3, we focus on Enschede-Noord, which is the neighbourhood of the campus. Here, the monthly fluctuations are more visible, which is logical due to a lower population. Lastly, in Figure 4 we see the number of assaults in the district Drienerveld-U.T., where the

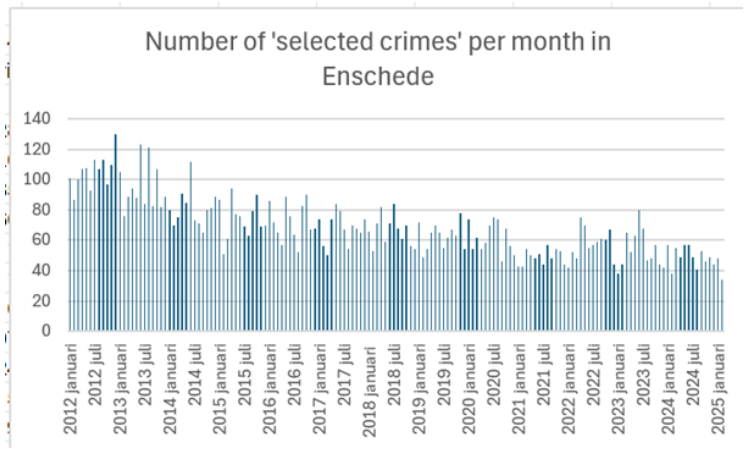

FIGURE 2: Number of assaults per month in Enschede.

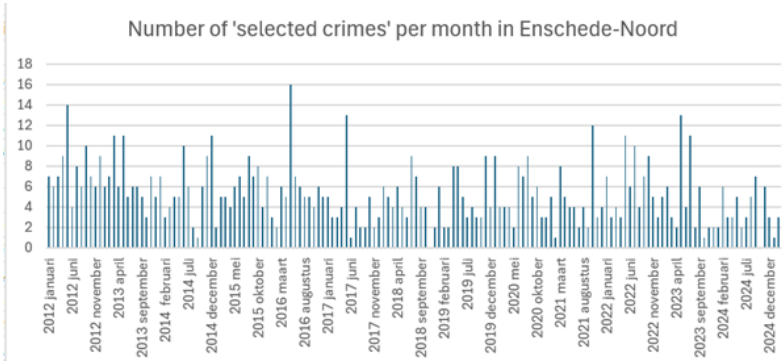

FIGURE 3: Number of assaults per month in Enschede-Noord.

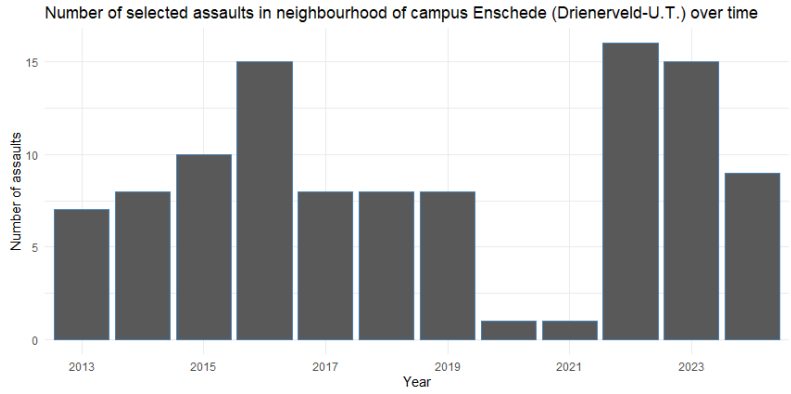

FIGURE 4: Number of assaults in district campus Enschede.

campus is located. One notable difference is that hardly any assaults have happened due to the COVID-19 outbreak, which was in 2020-2021. Note that this plot is only about very small numbers, in a range from 1-16.

These plots suggest that both long-term trends and small variations in the data should be considered in the model.

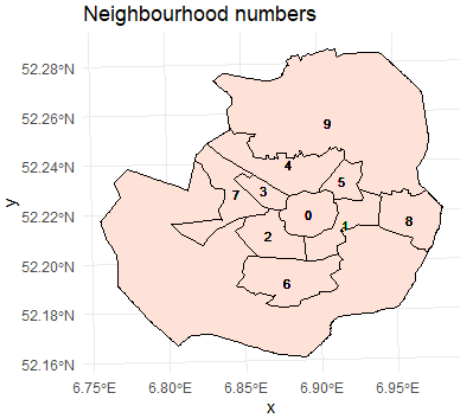

FIGURE 5: Numbering of neighbourhoods - Table 4

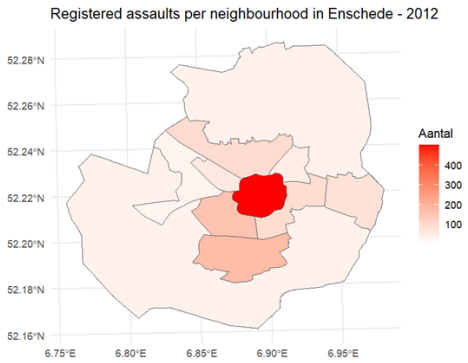

FIGURE 6: Number of assaults - 2012

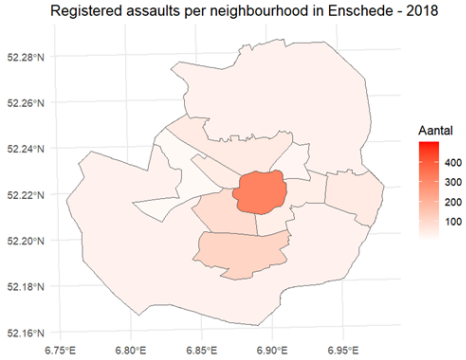

FIGURE 7: Number of assaults - 2018

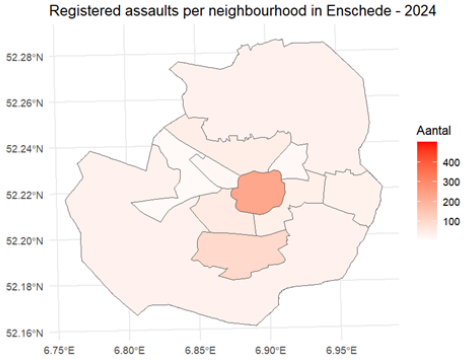

FIGURE 8: Number of assaults - 2024

2.4.2 Spatial trends

In addition to temporal variation, spatial effects are also present for these assault data. We start by giving a number to every neighbourhood, corresponding to the numbering of the CBS. This numbering can be found in Table 4 and with the corresponding map in Figure 5. The neighbourhood **Enschede-Noord** is in bold font, to express this is the area of interest in this paper.

Figure 6 shows that the highest number of assaults occur in the city centre. This could be due to higher population density, more activity, and greater foot traffic. In contrast, outlying neighbourhoods show lower number of assaults By comparing Figure 6 and Figure 8, we can state that the city centre shows the largest decline. Since the colour scale is consistent between the separate maps, the fading of the red area shows a big decrease. However throughout the whole area of Enschede, the colours have faded out. We will also analyse this statistically in Section 4. This is because the scale is identical, whereas the red centre has smoothened out. It could be that only the city centre is responsible for the decrease of assaults, which we saw in Section 2.4.1.

We also want to explore the dataset with the data about neighbourhoods. We do this to already find some candidates to answer subquestion 1, defined in Section 1. These variables are visualized in Figure 9, 10, 11 and 12.

Number neighbourhood	Name neighbourhood
0	Binnensingelgebied
1	Hogeland - Velve
2	Boswinkel - Stadsveld
3	Twekkelerveld - T.H.T.
4	**Enschede-Noord**
5	Ribbelt-Stokhorst
6	Enschede-Zuid
7	Bedrijfsterreinen Enschede-West
8	Glanerbrug en omgeving
9	Landelijk gebied en kernen

TABLE 4: Numbers of the map in Figure 5 with their names.

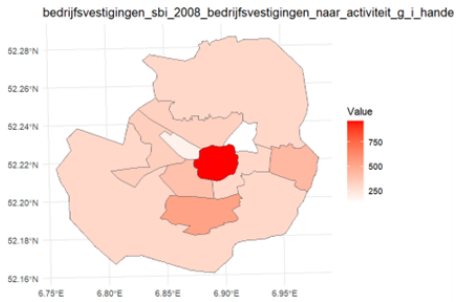

FIGURE 9: Number of eating and drinking facilities in 2024

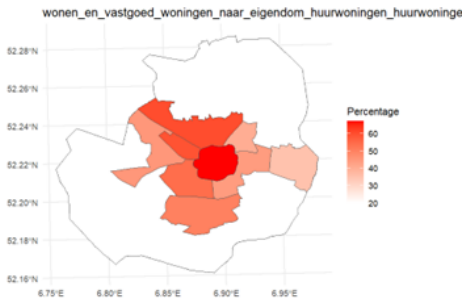

FIGURE 10: Number of rental houses in 2024

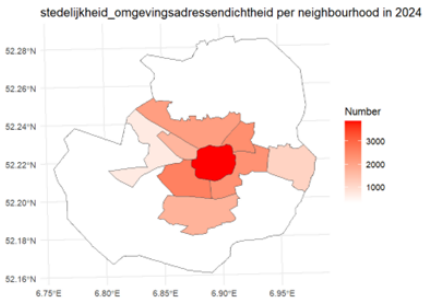

FIGURE 11: Address density in 2024

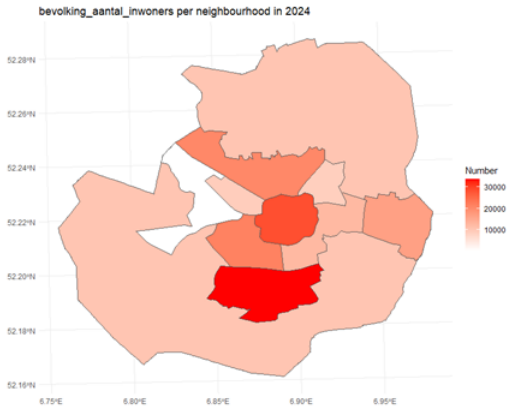

FIGURE 12: Number of inhabitants in 2024

3 Methodology

In the previous section, we did an exploratory data analysis and showed some visualizations to show potential spatial and temporal patterns in crime. In this section, we move from exploration to the actual model to answer the research questions formulated in Section 1. We want to build a model which explains the number of recorded crimes per neighbourhood

based on several explanatory variables. Given that the data is count data, we use a Poisson regression model. The remainder of this section will show the structure of the model and show which variables we will need in the model.

3.1 Poisson regression and Generalized Linear Models

To model the number of assaults, we use a Generalized Linear Model (GLM) with a Poisson distribution. A GLM extends a classical linear model, in the sense that it allows for response variables that follow a distribution other than the normal distribution. GLM's are very useful with non-continuous outcomes, such as count data.

Recall that a Poisson model is defined as follows: Let Y_i denote a random variable being the number of assaults for observation i, and $\mu_i > 0$ be the expected number of events. Then $Y_i \sim \text{Poisson}(\mu)$, where:

$$P(Y_i = y_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \tag{1}$$

with expected value $\mathbb{E}[Y_i] = \text{Var}[Y_i] = \mu_i$.

A GLM consists of three components:

- The distribution of Y_i, which is Poisson in this research.
- A linear predictor $\eta_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{x}_i$ is a vector of explanatory variables, and $\boldsymbol{\beta}$ is a vector with parameters.
- A link function g relating μ_i and η_i, such that $g(\mu_i) = \eta_i$.

More extended, a GLM is defined in terms of a set of independent random (response) variables $Y_1, ..., Y_N$, all drawn from the same type of distribution from the exponential family (Normal, Poisson or Binominal). Y_i is the number of assaults in a specific neighbourhood-month combination. This is count data, which makes we assume a Poisson model.

Any distribution in the exponential family can be written in the form:

$$f(y_i; \theta_i) = \exp\{y_i b(\theta_i) + c(\theta_i) + d(Y_i)\}, \tag{2}$$

where θ_i is the canonical parameter for observation i, and $b(\cdot), c(\cdot)$ and $d(\cdot)$ are known functions depending on the distribution. For example, the distribution function for the Poisson distribution is given by:

$$f(y_i; \theta_i) = \exp\{y_i \log \theta_i - \theta_i - \log y_i!\} \tag{3}$$

For every i, Y_i has the same distribution form. This makes that the joint probability density function can be stated as follows:

$$f(y_1, ..., y_N; \theta_1, ..., \theta_N) = \exp\left\{\sum_{i=1}^{N} y_i b(\theta_i) + c(\theta_i) + d(y_i)\right\}. \tag{4}$$

As θ_i can be different for every i, we aren't interested in θ_i that much. However, for the GLM, we want to know more about the parameters $\beta_1, ..., \beta_p \in \mathbb{R}$ $(p < N)$ in the following equation:

$$g(\mu_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{5}$$

where $\mu_i = \mathbb{E}(Y_i)$, $\boldsymbol{x}_i^T$ is a vector of explanatory variables and g is a monotone link function such that (5) holds.[5]

3.2 Model structure and hypothesis testing

We now want to relate the expected number of assaults μ_i to explanatory variables. Because the Poisson distribution only takes non-negative integer values, we apply a log link function, so $g(x) = \log x$. This function ensures the expected count remains positive. In the case of this research, we have data which are taken over neighbourhoods of different sizes. Therefore, we assume that each observation has an associated exposure term n_i. Exposure reflects the amount of opportunity for an event to occur, which can vary for different observations. For example, population size or surface area are common measures for the exposure. For the different models, we have to decide which of the two types of scaling we use.

The model can be described as follows:

$$\mathbb{E}(Y_i) = \mu_i = n_i e^{\beta_0 + \beta_1 x_1^i + \cdots + \beta_p x_p^i}, \tag{6}$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are unknown parameters and $x_i^1, \ldots, x_i^p$ are the covariates for observation i. Taking the logarithm of (6), we get [5]:

$$\log(\mu_i) = \log(n_i) + \beta_0 + \beta_1 x_1^i + \cdots + \beta_p x_p^i. \tag{7}$$

Now, this expression is linear in the parameters $\beta_0, \beta_1, \ldots, \beta_p$. Also, the term $\log(n_i)$ is in the predictor as a known offset. This gives a standard Poisson regression with a log-link function. This makes sure that the expected count is always non-negative and that the coefficients β_k include the effects of the different explanatory variables.

To test whether an explanatory variable significantly improves the model, use a likelihood ratio test. The likelihood ratio test compares two nested models to test whether the additional covariate(s) significantly improves the model. We do this by rewriting (7) as follows:

$$\log(\mu_i) = \log(n_i) + \beta_0 + \beta_1 x_1^i + \cdots + \beta_q x_q^i + \gamma_1 x_{q+1}^i + \cdots + \gamma_{p-q} x_p^i, \tag{8}$$

where:

- $\beta_i, i = 1, ..., q$ are the parameters of basic variables, which are not dependent on the variable under consideration.
- $\gamma_i, i = 1, ..., p-q$ are the parameters of variables which are dependent on the variable under consideration.
- Here, $x_i, i = 1, ..., q$ should be the variables which are not under consideration, while $x_i, i = q+1, ..., p$ are the other variables.

Following from this, we define two different models:

- Model M_1: Here, the following expression holds:

$$\log(\mu_i) = \log(n_i) + \beta_0 + \beta_1 x_1^i + \cdots + \beta_q x_q^i.$$

- Model M_2: The complete model; (8) holds here

Then we formulate the hypothesis test as follows:

$$\begin{cases} H_0 : \gamma_i = 0 \quad \forall i \in \{1, ..., p-q\} \text{(the covariate has no effect)} \\ H_1 : \exists i \in \{1, ..., p-q\} \text{ s.t. } \gamma_i \neq 0 \quad \text{(the covariate has a significant effect)} \end{cases}$$

Now, let $\mathcal{L}_1$ and $\mathcal{L}_2$ be the likelihood function of model M_1 and M_2, respectively. Then the test statistic is given by [6]:

$$\Lambda = -2\left(\log \mathcal{L}_1 - \log \mathcal{L}_1\right).$$

By Wilks' theorem [7], under the H_0, the test statistic Λ approximately follows a chi-squared distribution, so the following holds:

$$p \approx \mathbb{P}(\chi_k^2 \geq \Lambda),$$

where χ_k^2 is the chi-squared distribution with k degrees of freedom. This makes we can find the p-value using R with the following line:

```
anova(basic_model, advanced_model, test = "Chisq")
```

If the p-value is below the significance level of $\alpha = 0.05$, we reject the null hypothesis. We then can conclude that the covariate significantly improves the model. In Section 3.3, we will elaborate on the specific model we use for every question.

3.3 Research questions

In this section, we will formulate how the Poisson regression model, formulated in Section 3.2 is used to answer the three sub-questions in Section 1. For each of these sub-questions, we define a pair of nested models and formulate a hypothesis test based on the likelihood ration principle. Now we can assess whether a certain covariate significantly improves the model fit.

3.3.1 What factors influence the number of assaults?

To determine which socio-demographic factors are significantly associated with the number of assaults, we build multiple nested Poisson regression models. For each covariate of interest (e.g. population size, share of social housing), we compare a basic model without the covariate to an extended model that includes it, as described in Section 3.2.

- Model M_1 (Basic model):

  $$\log(\mu_{i,j}) = \log(n_{i,j}) + \beta_0,$$

  where $n_{i,j}$ is the surface area of neighbourhood i in year j (which we assumed was equal for all years).

- Model M_2 (Advanced model):

  $$\log(\mu_{i,j}) = \log(n_{i,j}) + \beta_0 + \gamma_1 \cdot x_1^{i,j},$$

  where $x_1^{i,j}$ is the variable of interest in neighbourhood i in year j.

Then the hypothesis test is formulated as follows:

$$\begin{cases} H_0 : \gamma_1 = 0 & \text{(the covariate has no effect)} \\ H_1 : \gamma_1 \neq 0 & \text{(the covariate has a significant effect)} \end{cases}$$

This hypothesis test can be examined using the anova test, like we showed in Section 3.2. This procedure is repeated for each covariate. In Section 4.1, the results are presented and discussed.

3.3.2 For the neighbourhood of Enschede-Noord, is there a big growth in number of assaults last year?

To investigate whether Enschede-Noord has experienced a significant increase in the number of assaults in 2024, we make a new binary indicator variable x_{2024}, which equals 1 for observations in 2024 and 0 otherwise. Then we again fit two models:

- Model M_1 (Model with covariates without indicator variable):

$$\log(\mu_{i,j}) = \log(n_{i,j}) + \beta_0 + \beta_1 x_1^{i,j} + ... + \beta_q x_q^{i,j},$$

  where $x_{i,j}$ are the covariates found by subquestion 1 in Section 4.1.

- Model M_2 (Model with covariates with indicator variable):

$$\log(\mu_{i,j}) = \log(n_{i,j}) + \beta_0 + \beta_1 x_1^{i,j} + ... + \beta_q x_q^{i,j} + \gamma_1 x_{2024},$$

  where γ_1 is the coefficient of the indicator variable.

From this, we can formulate a hypothesis test:

$$\begin{cases} H_0 : \gamma_1 = 0 & \text{(there is no outbreak in 2024)} \\ H_1 : \gamma_1 \neq 0 & \text{(there is an outbreak in 2024)} \end{cases}$$

If the test shows that the p-value is significant, we have shown that there is an unnatural rise in assaults in Enschede-Noord.

3.3.3 What can we say about the temporal trend of the assaults?

To assess whether there is a temporal trend in the assaults (overall decrease or increase), we include a covariate, which represents the year of observation.

- Model M_1 (Basic model without time):

$$\log(\mu_{i,j}) = \log(n_{i,j}) + \beta_0,$$

  where $n_{i,j}$ is the the surface of neighbourhood i.

- Model M_2 (Model fitted with time):

$$\log(\mu_{i,j}) = \log(n_{i,j}) + \beta_0 + \gamma_1 x_1^i,$$

  where $x_1^{i,j}$ is equal to the year j.

The corresponding hypothesis test is:

$$\begin{cases} H_0 : \gamma_1 = 0 & \text{(there is no time influence)} \\ H_1 : \gamma_1 \neq 0 & \text{(there is a time influence).} \end{cases}$$

Using the visualizations shown in Section 2.4.1 we can check if the number of assaults is structurally increasing or decreasing over time.

4 Results

In this section, we present the results of the statistical models and tests described in Section 3.2. We begin by examining which socio-demographic variables have a significant impact on the number of assaults. Once decided this, we use these variables in the further analysis, to determine if there is an outbreak in 2024, which was the actual reason this research is done. We will also formulate a model which can be used by other municipalities, to predict what will be the course of the number of assaults on the longer term. We will elaborate on that in the conclusion.

4.1 Influence of socio-demographic factors

Now we will check for factors that play a significant role in the number of assaults in Enschede. We do this by building a GLM (presented in Section 3.2) and testing if two different models are significantly different. In Table 5, the results of these tests are presented. We always fit the model on annual data of the 10 neighbourhoods.

x_i: $i=$	**Variable**	**Offset term**	**P-value**	**Deviance reduction**
1	Number of inhabitants	Surface area (km^2)	$2.2 * 10^{-16}$	11559
	log(Number of inhabitants)	Surface area (km^2)	$2.2 * 10^{-16}$	10593
2	Hospistality industry	Surface area (km^2)	$2.2 * 10^{-16}$	13192
3	Percentage households below social minimum	Surface area (km^2)	$2.2 * 10^{-16}$	13670
4	Rental housing	Surface area (km^2)	$2.2 * 10^{-16}$	20399
	Rate of urbanization	Surface area (km^2)	$2.2 * 10^{-16}$	22393
	Average value houses	Surface area (km^2)	$2.2 * 10^{-16}$	15914
	Hospitality industry * Number of inhabitants	Surface area (km^2)	$2.2 * 10^{-16}$	16905

TABLE 5: P-values of different covariates

Looking at Table 5, we see that every p-value lies far below our significance level of 0.05, which means we can certainly take these all into our model. There are more variables, which are significant. However we will not include all variables in the final model. There are several reasons for this. First, some variables are very much correlated, such as the number of inhabitants and the density of inhabitants. If we include all these variables, the model becomes very much more unstable and the model is more difficult to interpret. It could also be that we overfit the model, as we only fit our data on 120 data points.

From this, we can conclude that the above are all significantly important for our model. Therefore, we should use this information in our model.

If we take from all correlated variables the ones with the highest deviance reduction, we will get a model which has smaller error. This model can be described by the following equation:

$$\mu_{i,j} = \log(n_i) + \beta_0 + \beta_1 x_1^{i,j} + ... + \beta_4 x_4^{i,j}, \tag{9}$$

where x_i's are as indicated in Table 5. In Figure 13, the residuals can be seen. In this figure, we see that it keeps fluctuating quite a lot. However, we saw in Table 5 that all these variables are significantly useful. Therefore, we add some more variables: Average value of houses, and the rate of urbanization. In Figure 14, we can see the result of this. The

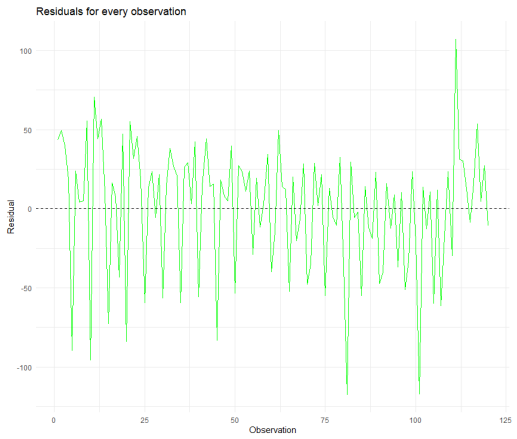

FIGURE 13: Residuals plotted for all 120 observations

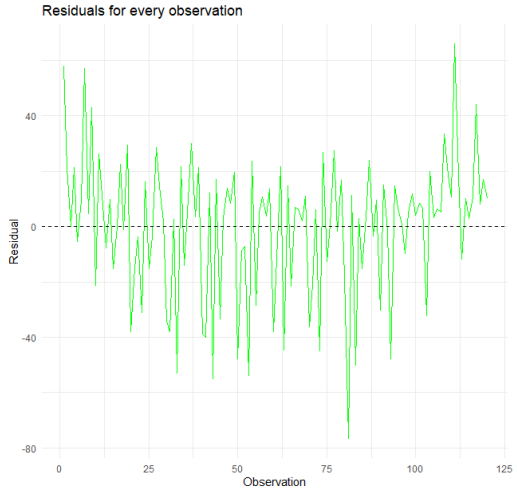

FIGURE 14: Residuals for 120 observation, variables added.

error has become some smaller. However, we still see some trend. This is to be improved in Section 4.3

4.2 Trends in Enschede-Noord

Executing the hypothesis test formulated in 3.3.2 with the model, we get a p-value of 0.090. This indicates we can not reject H_0, which makes we can statistically not say that there is an outbreak in Enschede-Noord from the data. Therefore, we don't take this indicator function into our model.

4.3 Temporal trends across Enschede

Executing the hypothesis test formulated in 3.3.3, we get a p-value of $2.2 * 10^{-16}$. This indicates we reject H_0, which makes we can statistically say that there is a shift over time in the number of occurrences. Looking to the visualisation, made in Figure 1, we see that this is actually a decrease in the number of assaults. Therefore, we also want to include time into our model. This results in Figure 15 We conclude with a reasonable model, as we stay around maximum of 25 deviation, while the original data ranged from 30-300. So

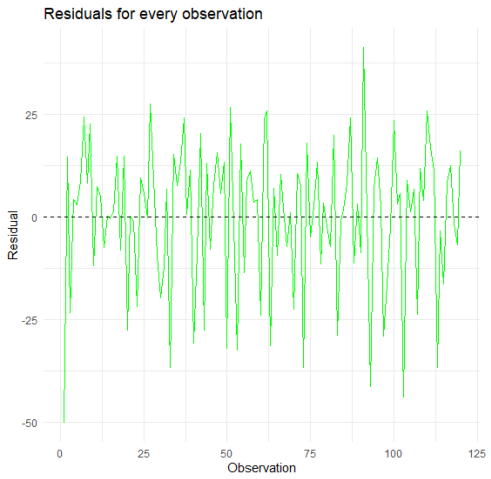

FIGURE 15: Residuals for 120 observation, time variable added.

this is a maximum deviation of 10%.

5 Conclusion and Discussion

In this thesis, we analysed the spatio-temporal patterns about the city and municipality of Enschede. For this we used crime data from 2013 up to and including 2024. Together with socio-demographic data, we fitted this to a Generalized Linear Model, specifically, Poisson regression model. From this research, we can conclude that the following factors significantly influence the number of public assaults:

- Number of inhabitants
- Hospitality industry
- Percentage of households below social minimum
- Rental housing
- Time
- Average value of houses
- Rate of urbanization

We did however not include all variables in the model, as some variables were too strongly related to each other.

Because of the series with assaults on the campus area, we wanted to know if there is a significant increase. By testing, we can say that we can not conclude that based on the data. This however does not mean there is no increase of such assaults. It could be very well that some cases are not reported to the police, which causes a lack of information. To discover these cases, one should do more research by asking people if they ever experienced such a assault. This also is a recommendation for some further research.

Lastly, we showed that there is some temporal factor in charge. By looking to the visualizations, we can conclude that this temporal factor results in a decrease of number of assaults in Enschede.

For the analysis annual data was available from the CBS for the explanatory variables. It would be very useful if there was monthly data available, which was the case for the crime dataset. This made that we lost quite some precision of this dataset. Also, we could not include seasonal effects, which are not visible on a scale of one year. Sometimes, the data with explanatory variables was not perfectly complete or consistent. We would have liked to insert data about migration and national backgrounds in the model. But CBS changed the definition of all these variables in 2016. This makes that this is not a consistent variable. We assumed for this research that Enschede is an isolated area, so no influence from outside. This is quite a rough assumption, especially as Enschede is located close to the border of Germany.

For further research, one could consider to use a negative binomial model, as this model can handle variation of data better.

To summarize, we built a model that takes both spatial and temporal patterns in assault data. Policymakers can make use of this model to investigate whether there will be an outbreak, or if their city is facing a growth in number of assaults. This can be useful, as policymakers can decide on preventing measures.

References

[1] *Unrest on campus after series of violent incidents*, Dec. 2024. [Online]. Available: https://www.utoday.nl/news/74981/unrest-on-campus-after-series-of-violent-incidents.

[2] *Data.politie.nl - geregistreerde misdrijven; soort misdrijf, wijk, buurt, maandcijfers.* [Online]. Available: https://data.politie.nl/#/Politie/nl/dataset/47022NED/table.

[3] *Wijk- en buurtkaart 2023.* [Online]. Available: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2023.

[4] *Kerncijfers wijken en buurten 2004-2024.* [Online]. Available: https://www.cbs.nl/nl-nl/reeksen/publicatie/kerncijfers-wijken-en-buurten.

[5] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, Third edition. Boca Raton, FL: CRC Press, 2008, ISBN: 978-1-58488-951-9.

[6] H.-p. Lai and C. J. Huang, "Likelihood ratio tests for model selection of stochastic frontier models," *Journal of Productivity Analysis*, vol. 34, no. 1, pp. 3–13, Aug. 2010, ISSN: 1573-0441. DOI: 10.1007/s11123-009-0160-8.

[7] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, Mar. 1938, Publisher: Institute of Mathematical Statistics, ISSN: 0003-4851, 2168-8990. DOI: 10.1214/aoms/1177732360. [Online]. Available: https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-9/issue-1/The-Large-Sample-Distribution-of-the-Likelihood-Ratio-for-Testing/10.1214/aoms/1177732360.full.

Appendix

A.1. R code for data preprocessing

```r
library(dplyr)
library(sf)
library(spatstat.geom)
library(ggplot2)
library(gganimate)

wijken_lijst = c(
  "Enschede",
  "Wijk 00 Binnensingelgebied",
  "Wijk 01 Hogeland - Velve",
  "Wijk 02 Boswinkel - Stadsveld",
  "Wijk 03 Twekkelerveld - T.H.T.",
  "Wijk 04 Enschede-Noord",
  "Wijk 05 Ribbelt - Stokhorst",
  "Wijk 06 Enschede-Zuid",
  "Wijk 07 Bedrijfsterreinen Enschede-West",
  "Wijk 08 Glanerbrug en omgeving",
  "Wijk 09 Landelijk gebied en kernen"
)

# load dataset
complete_dataset=read.csv("C:/Users/bartw/OneDrive - University of Twente/School/2024-2025/BA/Data/Misdrijven_per_wijk_en_buurt_per_maand_01052025_125907.csv",sep=";")
View(complete_dataset)
# sum over selected crime types and aggregate over months
jaarcijfers = filter(complete_dataset, Soort.misdrijf != "Totaal misdrijven", !(Wijken.en.buurten %in% wijken_lijst))
jaarcijfers$Perioden = substr(jaarcijfers$Perioden, 1, 4)
jaarcijfers =  summarise(group_by(jaarcijfers, Perioden, Wijken.en.buurten), Totaal = sum(Geregistreerde.misdrijven..aantal., na.rm = TRUE), .groups = "drop")
#write.csv(jaarcijfers, file = "C:/Users/bartw/OneDrive - University of Twente/School/2024-2025/BA/jaarcijfers.csv", row.names = FALSE)
```

A.2. R code for preprocessing model fitting

```r
library(dplyr)
library(janitor)
library(tibble)
library(ggplot2)
library(sf)

path = "C:/Users/bartw/OneDrive - University of Twente/School/2024-2025/BA/Data/Kerncijfers"
years = 2013:2024
all_data = list()

for (year in years){
```

```
  current_file = file.path(path, paste0("Kerncijfers_wijken_en_
    buurten_",year,".csv"))
  cijfers = read.csv(current_file, sep = ";",skip=3, check.names =
    FALSE, stringsAsFactors = FALSE)
  cijfers = data.frame(t(cijfers))
  colnames(cijfers) = as.character(unlist(cijfers[1,]))
  cijfers = cijfers[-c(1,2),]
  cijfers = rownames_to_column(cijfers, var = "WijkNaam")
  cijfers = clean_names(cijfers)
  cijfers = filter(cijfers, regioaanduiding_soort_regio == "Wijk")
  cijfers$year = year
  all_data[[as.character(year)]] = cijfers
}

all_data = bind_rows(all_data)
all_data = all_data[, c("year", setdiff(names(all_data), "year"))]
colnames(all_data)
colnames(all_data)[colSums(is.na(all_data)) == 0]

jaarcijfers = read.csv("C:/Users/bartw/OneDrive - University of
    Twente/School/2024-2025/BA/jaarcijfers.csv")
View(all_data)
cols_no_na <- names(all_data)[colSums(is.na(all_data)) == 0]
fit_data = all_data[, c("year", "wijk_naam", "bevolking_aantal_
    inwoners", "bedrijfsvestigingen_sbi_2008_bedrijfsvestigingen_
    naar_activiteit_g_i_handel_en_horeca","inkomen_inkomen_van_
    huishoudens_huish_onder_of_rond_sociaal_minimum")]
fit_data$huurwoningen = coalesce(all_data$wonen_woningen_naar_
    eigendom_huurwoningen_huurwoningen_totaal, all_data$wonen_en_
    vastgoed_woningen_naar_eigendom_huurwoningen_huurwoningen_totaal
    )
fit_data$woz = coalesce(all_data$wonen_en_vastgoed_gemiddelde_woz_
    waarde_van_woningen, all_data$wonen_gemiddelde_woz_waarde_van_
    woningen, all_data$wonen_gemiddelde_woningwaarde)
fit_data = left_join(jaarcijfers, fit_data, by = c("Perioden"="year
    ", "Wijken.en.buurten"="wijk_naam"))
fit_data = filter(fit_data, Perioden %in% years)
fit_data$Perioden = factor(fit_data$Perioden)
numeric = c("Totaal", "bevolking_aantal_inwoners","Perioden", "
    bedrijfsvestigingen_sbi_2008_bedrijfsvestigingen_naar_activiteit
    _g_i_handel_en_horeca", "inkomen_inkomen_van_huishoudens_huish_
    onder_of_rond_sociaal_minimum","huurwoningen", "woz")
fit_data[numeric] = lapply(fit_data[numeric],
                           function(x) {
                             val = as.numeric(gsub(",",".",x))
                             ifelse(is.na(val),0,val)
                           }
)

data_enschede_noord = filter(fit_data, Wijken.en.buurten == "Wijk
    04 Enschede-Noord")
data_enschede_noord$y2024 = ifelse(data_enschede_noord$Perioden
    ==2024,1,0)
View(data_enschede_noord)

```

```r
wijken = st_read("C:/Users/bartw/OneDrive - University of Twente/School/2024-2025/BA/WijkBuurtkaart_2023_v2/WijkBuurtkaart_2023_v2/wijkenbuurten_2023_v2.gpkg", layer="wijken")
enschede_wijken = filter(wijken, gemeentenaam == "Enschede")
oppervlaktes = tibble(Wijken_en_buurten = enschede_wijken$wijknaam, oppervlakte_km2 = as.numeric(st_area(enschede_wijken))/1e6)
fit_data = left_join(fit_data, oppervlaktes, by = c("Wijken.en.buurten"= "Wijken_en_buurten"))
fit_data = left_join(fit_data, select(enschede_wijken, wijknaam, stedelijkheid_adressen_per_km2), by = c("Wijken.en.buurten"="wijknaam"))
fit_data$log_oppervlakte_km2 = log(fit_data$oppervlakte_km2)

model1 = glm(
  Totaal ~  Perioden + bevolking_aantal_inwoners +
    bedrijfsvestigingen_sbi_2008_bedrijfsvestigingen_naar_activiteit_g_i_handel_en_horeca +
    inkomen_inkomen_van_huishoudens_huish_onder_of_rond_sociaal_minimum +
    huurwoningen +
    woz+
    stedelijkheid_adressen_per_km2,
  data = fit_data,
  family = poisson(link = "log"),
  offset = log(oppervlakte_km2)
)

model_basis = glm(
  Totaal ~  1,
  data = fit_data,
  family = poisson(link = "log"),
  offset = log(oppervlakte_km2)
)

model_extended = glm(
  Totaal ~  bevolking_aantal_inwoners *
    bedrijfsvestigingen_sbi_2008_bedrijfsvestigingen_naar_activiteit_g_i_handel_en_horeca,
  data = fit_data,
  family = poisson(link = "log"),
  offset = log(oppervlakte_km2)
)

summary(model_basis)
data_enschede_noord
anova(model_basis,model_extended,test = "Chisq")

fit_data$voorspeld = predict(model1, type = "response")
fit_data$Pearson_res = (fit_data$Totaal - fit_data$voorspeld)/sqrt(fit_data$voorspeld)
fit_data$residu = fit_data$Totaal - fit_data$voorspeld
sum_of_squares = sum(fit_data$Pearson_res^2)
sum_of_squares
ggplot(fit_data, aes(x=1:nrow(fit_data))) +
  geom_line(aes(y = residu), color = "green") +
  geom_hline(yintercept = 0, linetype = "dashed") +
```

```
  labs(x = "Observation", y= "Residual", title = "Residuals for
      every observation") +
  theme_minimal()

ggplot(fit_data, aes(x = Totaal, y = voorspeld)) +
  geom_point() +
  geom_abline(slope = 1, intercept = 0, linetype = "dashed", color
      = "red") +
  labs(
    x = "Echte waarde (Totaal)",
    y = "Voorspelde waarde",
    title = "Voorspeld vs. echt aantal misdrijven"
  ) +
  theme_minimal()
```