Utrecht University

Master Thesis Artificial Intelligence

Global models and local patterns for crime prediction from weather data

Jurian Baas

supervisors

Dr. Mirna v. Hoek - KNMI
Dr. Andrea Pagani - KNMI

examinators

Dr. Ad Feelders - Utrecht University
Prof. Arno Siebes - Utrecht University

September 14, 2017

Acknowledgements

This work would not have been possible without the guidance of my supervisors.

I would therefore like to thank my university supervisor Dr. Ad Feelders for his guidance, creative ideas, insights and willingness to talk in depth about the subject field almost every week.

Of course I would like to thank my KNMI supervisor Dr. Mirna van Hoek, for settling me in, helping me with orienting myself within the KNMI and her always helpful and timely feedback during the course of the project.

Additionally my thanks go out to Dr. Andrea Pagani and Dr. Raymond Sluiter from the KNMI and Drs. Frank Willemsen of the WODC for helping me with the technical issues such as providing the necessary data, giving feedback and suggesting helpful ideas.

Abstract

Weather influences people indirectly in many ways. Does this include criminal behavior? Previous research has shown a definite relationship between high temperatures and an increase in violent behavior. This thesis attempts to determine and examine more complex relationships with the use of both local and global machine-learning and data-mining models. The resulting global models perform only marginally better than simple baseline models. However, local patterns built with the Patient Rule Induction Method yield interesting subgroups that are in line with preceding research elsewhere.

Contents

Chapter 1

Introduction

1.1 Research motivation

We all have had the experience of postponing tasks because it was raining outside. The weather influences us in all sorts of ways, both psychologically and (therefore) behaviorally, such as making us feel more energetic, to go outside or rather stay inside. With this in mind, are criminals (and their victims) constrained by the same factors in their activities?

Before we can think of issuing warnings to the police about crime inducing weather conditions there needs to be an established correlation between them, and if one exists, exactly how and how strongly this correlation manifests itself. The relation between weather and crime can only be indirect, so any possible relation will most likely be subtle and subject to a great amount of noise. There exist two models that try to explain the cause of crime; the *interactional-* and *routine activity* hypotheses.

According to the interactional hypothesis people have their own way of dealing with (environmental) stress. Weather can be considered another stress-factor. When it is added to the already existing daily stress it can be the final straw that pushes an individual "over the edge" and then to commit a crime [1].

The routine activity hypothesis states that certain changes in weather lead to changes in our daily routines [2]. For instance, in the summer when the temperature is high, people tend to spend more time outside. Having more individuals outside means more potential victims. A larger number of open windows due to excess heat means more opportunity for a break in, etc.

According to these hypotheses there could be a relation between (certain) crimes and weather types, but it is not known what that relation is. Different kinds of crimes might depend on a variety of weather variables. We want to find whether, and if so, how crimes depend on different weather variables.

1.2 Problem Definition

The goal is to establish whether and how the number of crimes in a crime category depends on certain weather variables. Firstly, by determining if a statistical relationship between crime and weather can be ascertained. Secondly, by building predictive models and testing them on their ability to generalize to unseen data. Familiar techniques such as linear regression will be used to determine the existence of a possible relationship between a particular crime category and one or more weather variables.

However, other machine learning techniques could prove a better way to examine this hypothetical dependency. These considerations broaden the scope to also examine, research and possibly apply alternatives to linear regression such as Poisson regression models and support vector machines. Next to global models we shall also make use of a local model, namely the patient rule induction method. This work will also attempt to briefly explain the algorithms that are used as to provide additional context to what operations are being performed.

1.3 Methodology

1.3.1 Project Life-cycle

Data mining is a very creative process which requires a number of different (domain specific) skills to succeed. During the investigations it is often necessary to go back to a previous step to redo or refine a specific element. For this reason the **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining (CRISP-DM) was used in this project.

CRISP-DM is organized as a number of phases, each consisting of several generic tasks. This allows for the possibility to try out different data mining and machine learning techniques in a relatively consistent manner [3].

Figure 1.1 shows the general structure of the CRISP-DM cycle. Notice how the evaluation step does not always lead to deployment but often to a new cycle of data understanding and modeling. In practice it turned out that most time was spent going between the data preparation and modeling phases. New insights into the data are fed back into the modeling process to further refine and understand the underlying data.

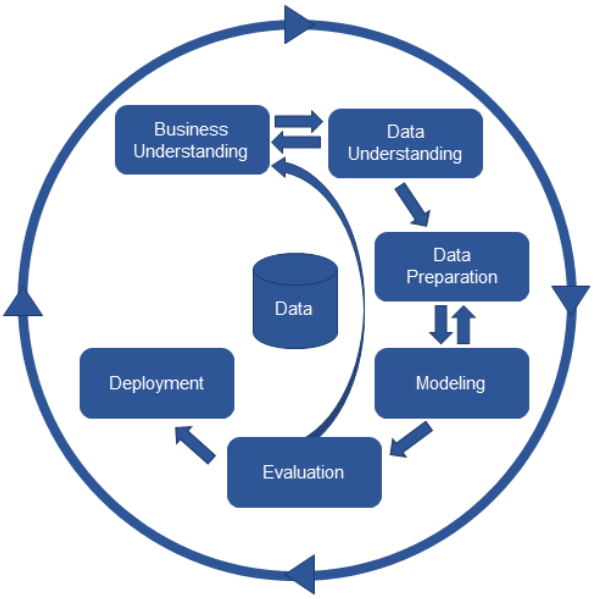


Figure 1.1: CRISP-DM cycle

1. **Business Understanding** This phase focuses on understanding the project from a business perspective. The objectives and requirements are used to construct a data mining problem definition. This project is less focused on the business side as the research is preliminary and has a strong exploratory character.

2. **Data Understanding** By becoming more familiar with the data it is possible to extract hidden information and interesting subsets. Also problems in quality such as missing values are determined here. In the case of this project much of the data was collected and partly processed.

3. **Data Preparation** The final data sets that will be fed into the models are constructed in this phase. This often requires merging, transforming, sorting, selecting subsets and further cleaning. This task is usually performed many times over as new insights come available.

4. **Modeling** Various modeling techniques are used in the modeling phase to generate new and interesting insights into the data. Often this requires to go back to the data preparation phase in order to reformat the data so it can be applied for some specific modeling technique. This phase should give models that are high quality from a data mining perspective, e.g. they avoid things like overfitting.

5. **Evaluation** When the modeling phase generates interesting models that seem to fulfill the business objectives they have to be critically evaluated. This means checking whether some objective has been left out or not fully considered.

6. **Deployment** The knowledge gained in the previous phases has to be organized in a way that is usable for the customer. This can range from writing a report to a fully implemented business wide application. This project does not include the final deployment phase in its scope except for the final report and the open source R package.

1.3.2 Software

The R programming language

The nature of this project requires a programming language that is "natural" to use for applied statistics techniques. *R* gives a solid base for using and building complex statistical models [4]. Many functionalities are already available via the *Comprehensive R Archive Network* (CRAN) and don't have to be developed from scratch. On the other hand, the software on CRAN was developed by third party contributors and is not in all cases reliable or well tested. Therefore the packages that were used in this work are those that have proven themselves as dependable by their overwhelming usage in the R community.

Rstudio

The development environment for R that comes to mind first is *Rstudio*. Created as an open-source initiative it is well suited for this project. It supports searching and downloading packages and dependencies from CRAN as well as the framework to develop your own package if necessary. This functionality was welcome as part of this work a new R package was developed and released on CRAN.

1.3.3 Global and local models

The main difference between global and local models is that global models try to approximate a function that describes all the data best, while a local model will attempt to find a function that only characterizes a (small) subset of the data. Therefore a global model can be used to make predictions on the entire feature space of the data.

The local model can give us a description of its boundaries and a prediction value based on the quality function (explained in chapter 4) used to construct it. Another advantage is that during construction the decisions made to determine the boundaries of a box are not constrained by having to take into account the observations that fall outside the box. It does, as previously stated, only describe a proportion of the data. This deficiency with local models can be somewhat reduced by building many local models, each describing a different (overlapping) part of the data.

1.4 Related work

The established research shows that there are indeed relationships between weather phenomena and violence. Certain ambient environmental features such as temperature (and even smell) can have a negative impact on subject aggression. Laboratory research has shown that high ambient temperatures can both increase and decrease aggression [5, 6]. This is usually interpreted as a curvilinear relationship between temperature and aggression, i.e. as temperature increases so does aggression up to a point, after which aggression starts to decrease. These laboratory conditions are somewhat suspect however, as a hot temperature is a salient and unusual condition. The subject might infer from this the experiment's intent and work against the supervisors in some effort of defiance. Although people often underestimate the effects of situational factors, which may include temperature [7]. An additional difficulty for laboratory experiments that it is hard to study criminal behavior in a laboratory setting where people know they are being monitored.

It is of course trivial to postulate that a curvilinear relationship *must* exist because there always is a point where it becomes too hot for humans to stay alive. The question is rather where this inflection point occurs and if these conditions actually take place in the outside world.

Field studies have also been performed, usually in the United States. In one instance, several riots were examined on the environmental conditions when and where they occurred. Here a definite curvilinear relationship was found, the frequency of riots increased with temperature up to around 30°C, then decreased as temperature rose more. However these results do not take into account the interdependency of the variables, as there were fewer riots on very hot days not because they were particularly uncomfortable but rather more likely because there are fewer of such days [8]. Other examinations of assault and temperature consistently found a positive relationship between the two [9, 10, 11, 12, 13].

Notably there was no relation found for robbery and temperature, perhaps due to the fact that robbery is more motivated by economic need and not as much by aggression [5]. It will be interesting to learn if the same pattern emerges in the Dutch weather and crime data.

With respect to seasonality there is a definite established relationship as well. As will be shown in the exploratory analysis of this work (section 2.1.2), the Dutch crime data suggests a similar pattern as was previously found. This pattern is that violent crime rates peak during the summer months, while property crime rates peak during the winter [14, 15].

In the vast majority of cases some sort of ordinary least-squares (OLS) model was used to establish the relationships between weather conditions and crime rates. This is not always appropriate when the rates for many of the units must be computed from small numbers of events. As population decreases, a crime rate of zero will be observed for a larger proportion of cases. So there is censoring at zero, depending on sample size [16]. In this thesis we also attempt to solve this problem using Poisson based regression methods, including a more complex two-step "zero inflation" approach.

If indeed these findings truly exist in the real world, ongoing processes like climate change could have a large impact on the number of violent crimes. As temperature increases worldwide, so would the number of non-economic crimes (perhaps also the amount of economically motivated crimes) [17]; with great (financial) cost to society. Therefore understanding these relationships could not only reduce crime in the short term, but also provide a better way to mitigate the rise of crime in the long term.

Chapter 2

Exploratory Analysis

2.1 Crime Data

The crime related data was provided by the WODC. (Wetenschappelijk Onderzoek- en Documentatiecentrum) This government organization is dedicated to doing research and collecting statistical information on the topic of security and justice. It also cooperates with external parties such as universities and research institutes.

2.1.1 Data Structure

The raw data comes in the form of 4.924.793 separate incidents in the period 2006 - 2015 (inclusive). All instances have been through the judicial process, this means that we have no information on the number of *reported* incidents. The actual number of instances could therefore be much higher than reflected by the data. The information available per crime instance is: the date, where the crime occurred (municipality), a description of the crime (category) and the age, gender, nationality and country of origin of the person who committed the crime.

Crime Categories

577 Different crime categories are contained by the raw crime data. These are often too specific for our purposes because we need a sufficient number of instances per day per category. Sometimes the same category is mentioned twice, once starting with a capital letter and once without. At other times a specific category is made for the municipality of Almelo. In order to have enough information per category on a daily basis they have been combined into 12 new super-categories:

Assault	Burglary	Discrimination	Domestic Violence
General	Hard Drug	Larceny	Murder
Rape	Robbery	Soft Drug	Vehicle Theft

Table 2.1: Combined crime categories

Many categories such as those associated with 'white collar' financially motivated crime like fraud, Ponzi-schemes and bribery have been excluded from the analysis. They most likely are not or only minimally influenced by weather or occur over a longer time period where the notion of weather becomes meaningless.[1]

Zero-day Data

Depending on the municipality and crime category, many - or even most - days do not have any crimes associated with them. This can cause some machine learning algorithms like linear regression to perform worse when taking these days into account. Likewise it can be regarded as 'unfair' to simply ignore them; As the weather conditions on those days could have been precisely those that contributed to the low number in the first place!

[1] Please see table 8.1 in the appendix for a full overview of category allocation.

We therefore include days with no crime in our analysis; and have to use methods like *zero-inflated models* in order to deal with excess zero's.

Aggregating by date and municipality

Further pre-processing was necessary in order to be able to merge the crime data with the available weather data. Therefore we group the instances by municipality and date:

Date	Municipality	Crime type	Crime count
01-01-2006	Amsterdam	Robbery	4
01-01-2006	Rotterdam	Assault	7
	⋮		
30-12-2015	Utrecht	General	11
31-12-2015	Den Haag	Larceny	11

Table 2.2: Crimes aggregated by municipality and date

Further analysis has been done using this aggregated form of the data.

2.1.2 Quirks and Anomalies

Long term crime trends

From figure 2.1 we can see that crime in the 12 categories listed in table 2.2 tends to decrease over time until around 2010. Afterwards it remains stable for a few years and then increases again. The final downwards trend in 2015 is due to cases still in due process and are therefore absent from the data. Note again that this trend might not reflect an *actual* decrease in crime in society in general as we can only report those cases that have been through the judicial process.

Missing values in the final period are especially noticeable among crime categories which typically take longer to solve, such as murder.

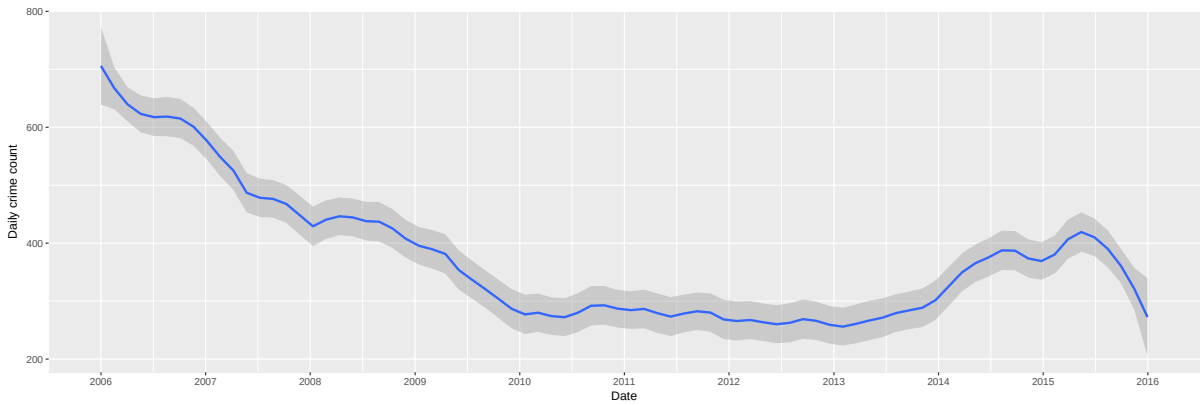


Figure 2.1: National summed daily crime

When we look more closely at the contribution of several specific crime categories in figures 2.4 and 8.1, it becomes clear that they can affect the national trend dramatically. The distributions of these sub-trends in turn are likely very much influenced by the policy of the Police and the Department of Justice. Examples of such policy changes could be cracking down on previously tolerated crimes and changing definitions and moving numbers under a new header.

Periods of high variance

The high crime numbers in the period 2006 are not caused by more daily crime in general, instead there are a number of 'spikes' - days with a very large number of crimes - on a certain location that together contribute to a higher average. It is unknown and unlikely that these days reflect an actual very temporary increase. More likely is some quirk in documentation and administration that caused these anomalies. Figure 2.2 shows how the variance dramatically increases in the year 2006:

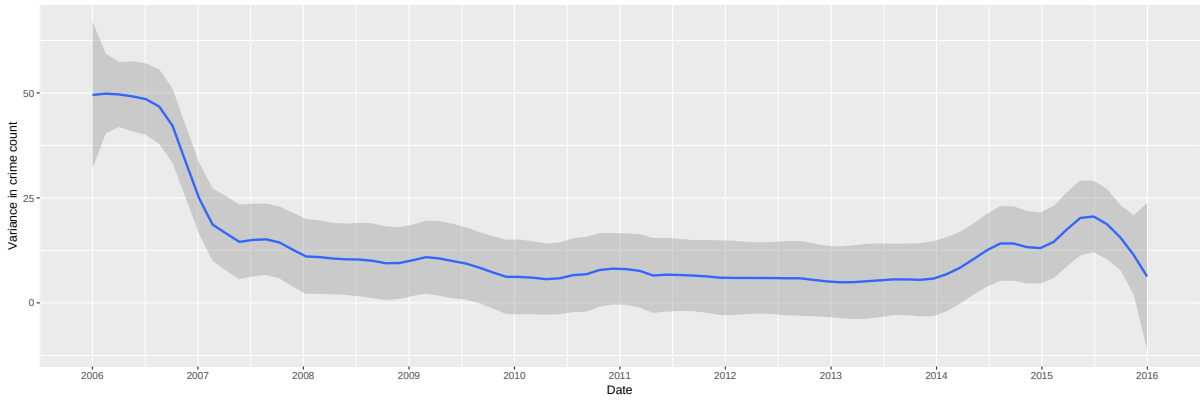


Figure 2.2: Variance in daily crime

Seasonal crime trends

Certain types of crime like robbery show a strong seasonal trend. This is in line with what has been found in previous research done in Chicago [6]. Figure 2.3 shows the average number of robberies per day over time. Peaks in the number of robberies seem to fall in the late autumn and early winter. Depending on the location this trend is largely maintained, although the years 2014 and 2015 seem to break this trend. But as we shall see later, when looking more closely at a single municipality such as Amsterdam there are also different trend-breaking years.

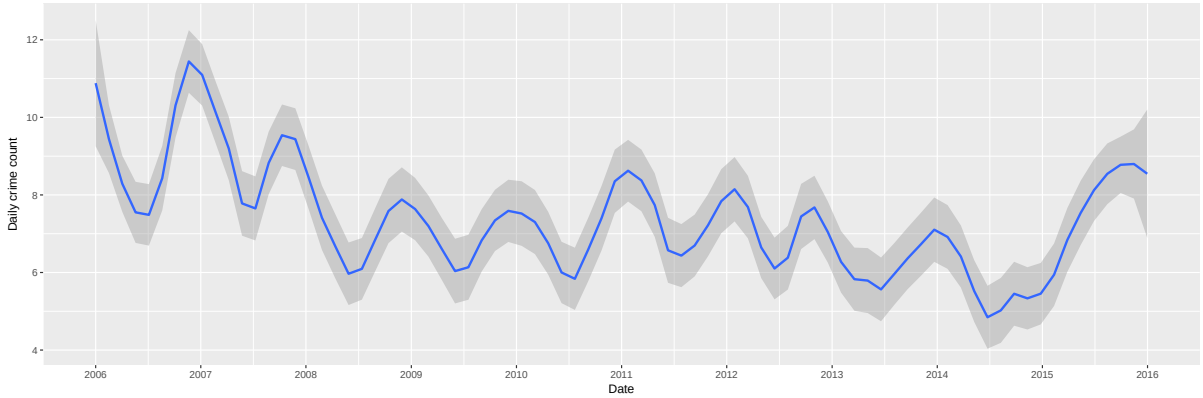


Figure 2.3: National average robbery count per day

When we look at other crime categories in figure 2.4 they sometimes follow a similar seasonal trend, though with peaks at different times. Additionally other types of crime are subject to high variability among certain years. Notice the dramatic increase in the number of soft drug related convictions starting in 2013 and the sharp drop-off in the number of murders in 2015. These are most likely caused respectively by a change in policy and a backlog of unsolved cases. Interestingly, many crime categories show a downwards trend.

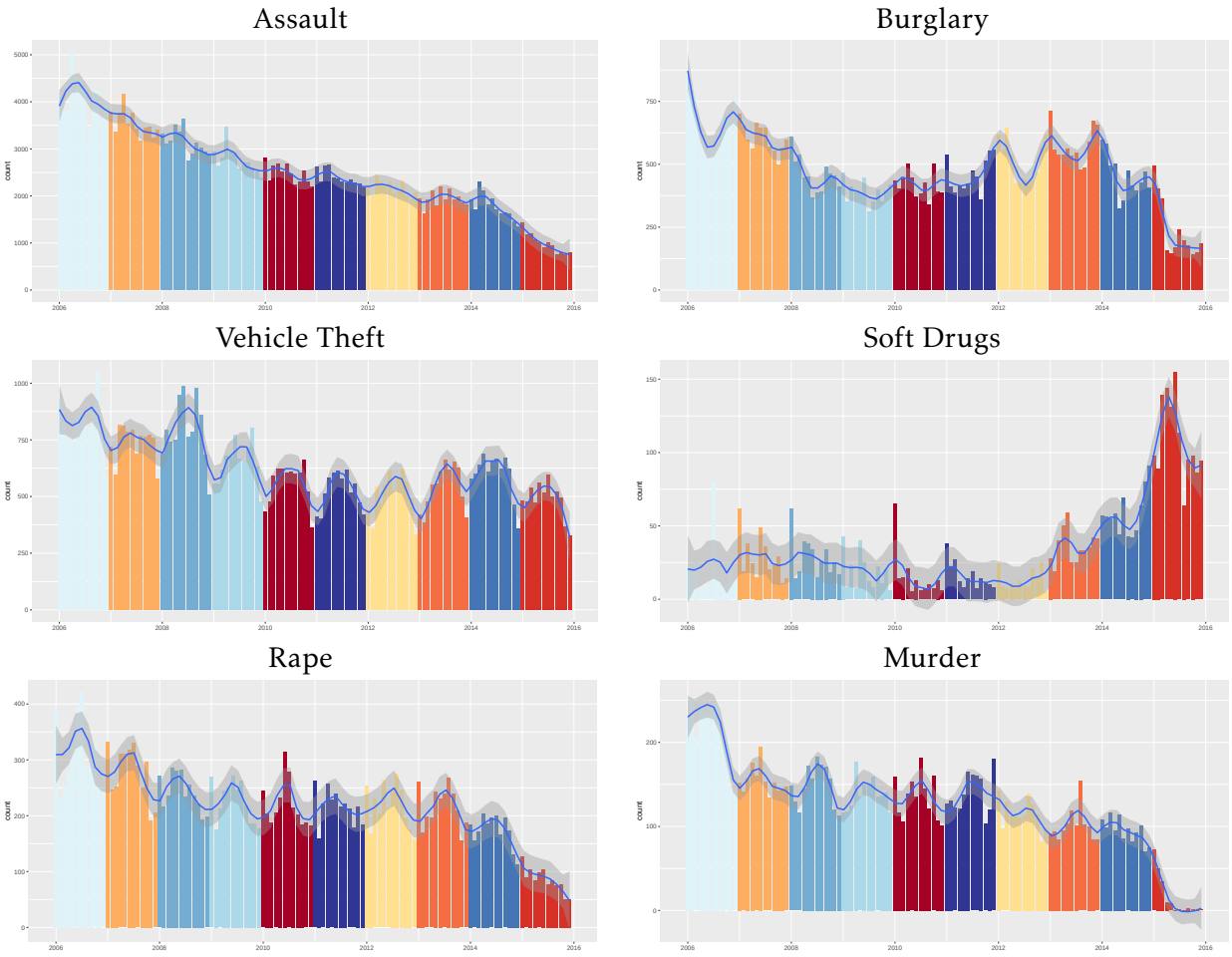


Figure 2.4: Monthly crime counts per category

Please consult figure 8.1 in the appendix for the distributions of the other crime categories.

Holidays and celebrations

Certain days that tend to coincide with partying and heavy drinking such as January 1^{st} (New Year's Day), April 30^{th} (Queensday) and April 27^{th} (Kingsday) show a strong increase in the number of violence related crimes. In figure 2.5 we can see that the number of assaults rises dramatically compared to the rest of the year on these particular days. The inclusion of these days in the data could potentially influence the results of the analysis. Giving importance to certain predictors that happen to be low or high during the celebrations.

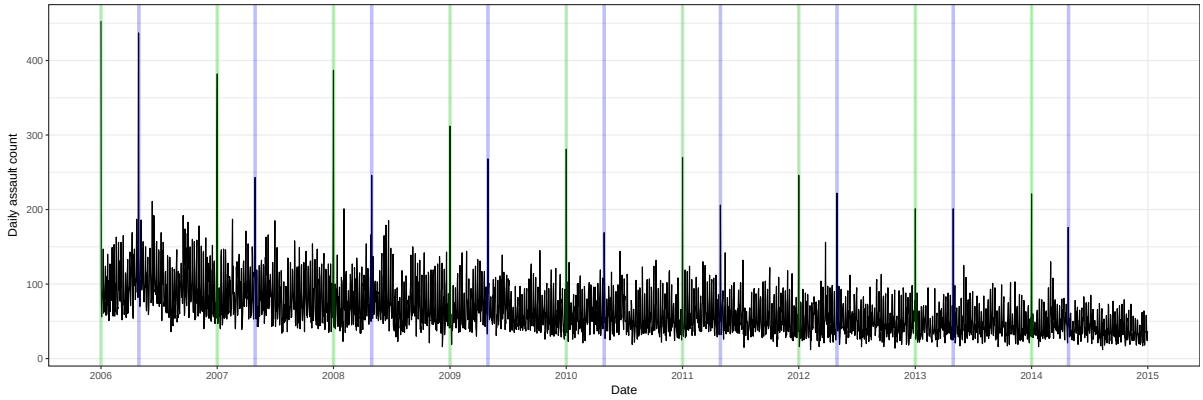


Figure 2.5: National daily number of assaults, New Year's Day highlighted in green, Queen- and Kingsday in blue

2.1.3 Weekday crime distributions

Certain crimes tend to happen more often during the weekend. Others like larceny show a notable drop on Sunday. These effects could influence the results when looking at weather parameters only. What if we predict high larceny but it just happens to be a Sunday too? These societal effects are most likely stronger than those from the weather, unless it is particularly extreme. [2]

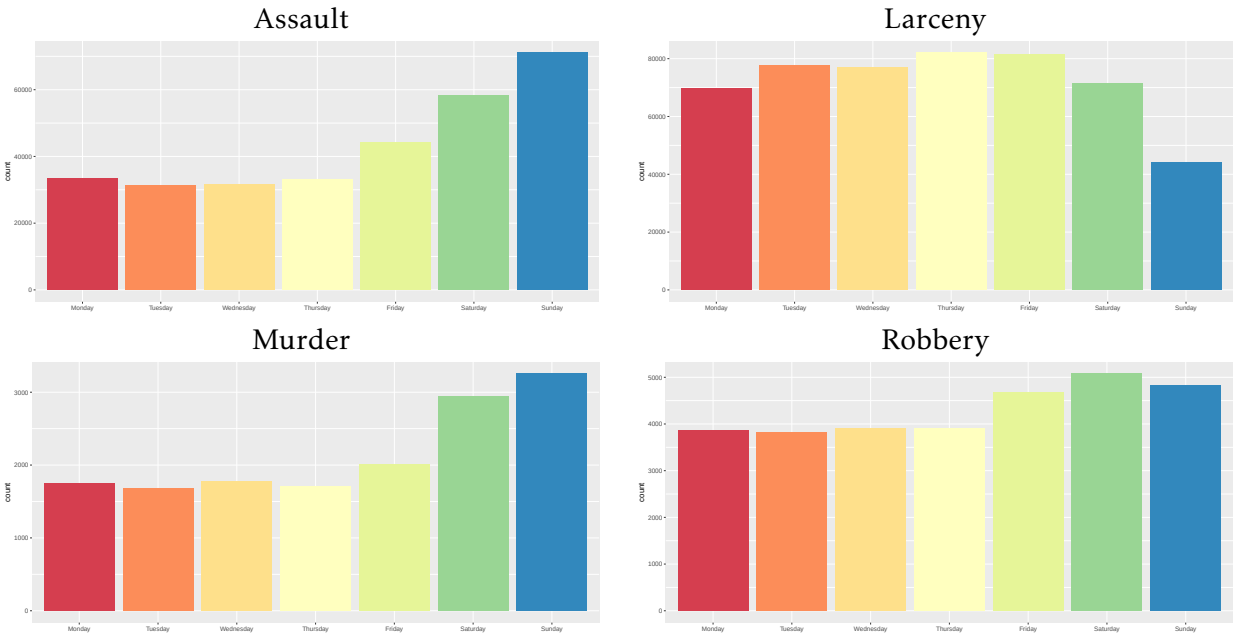


Figure 2.6: Weekday crime counts per category

[2] Please refer to figure 8.2 in the appendix for the weekday distributions of the other categories.

2.1.4 Compensating for population size

When comparing different municipalities it is useful to take the population size into account. For this reason population data was obtained from the Dutch National Bureau of Statistics (CBS) and combined with the crime data from the WODC. Difficulties in this approach were that both parties have slightly different naming conventions for certain municipalities (such as adding a period after an abbreviation of the province) and the fact that some municipalities have been merged in the period of 2006-2015. These do not appear separately in the CBS data but only in the years after they have been merged.

We calculate the crime density of a municipality by taking the total number of crimes committed and dividing it by the average population in that municipality over the ten year time period. Using the sets of all municipalities M, all populations for each municipality P and all days D we calculate the total crime per municipality C_m, the crime density D_m and the fraction of the municipality with the largest density F_m:

$$C_m = \sum_{d \in D} c_{m,d} \tag{2.1}$$

$$D_m = \frac{C_m}{P_m} \tag{2.2}$$

$$F_m = \frac{D_m}{\max_{m \in M} D_m} \tag{2.3}$$

We can visualize the 'crime density' per person as a ranking in figure 2.7. The municipality with the highest crime-density is placed at 1 (in this case Amsterdam) and every other municipality is ranked as a percentage of the maximum. When compensating like this, certain cities such as Amsterdam, Rotterdam and The Hague stay at their original rank while Utrecht drops from fourth to sixteenth place.

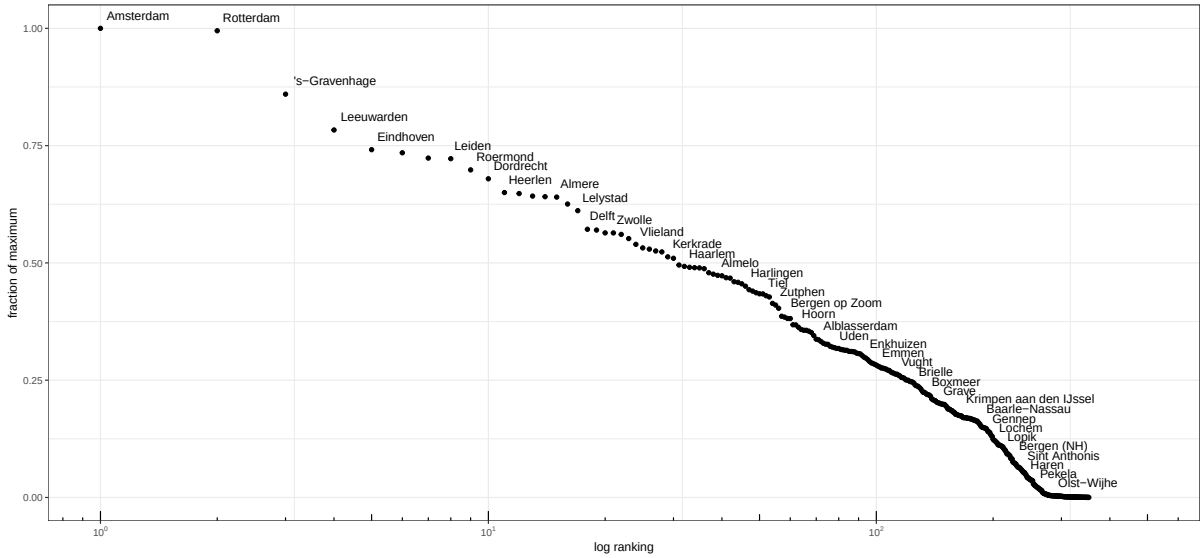


Figure 2.7: Municipalities ranked by crime density

These adjustments will be needed in future work to adjust for population size when municipalities other than Amsterdam are examined and compared.

2.2 Weather Data

The KNMI has provided weather data from (almost) every day in period 2006 - 2015 from 32 high end weather stations. Interpolation is used in order to get the weather information from all the municipalities that occur in the crime data. Methods used are mostly *Kriging* and *Thin Plate Splines* [18].

2.2.1 Data Structure

The weather data units and ranges are given in table 2.3. Some of these units are in very different scales, such as average temperature which has a range of (-14.79, 28.63) °C and radiation with a range of (0, 3160.1) Watts per m^2.

Predictor	Unit	Min value	Max value
Temperature	Degrees Celsius	-14.79	28.63
Relative Humidity	Percentage	31.26	101.04
Radiation	Watts per m^2	0	3160.1
Precipitation	Millimeters	0	137.22
Pressure	Millibar	972.6	1046.4
Sunshine	Hours	0	17.209
Windspeed	Meters per second	0.41	17.379
Wind Direction	Degrees	0	360

Table 2.3: Weather data units and ranges

2.2.2 Centering and Scaling

Some methods of analysis such as linear regression and support vector machines can be sensitive to these kind of differences in scale. In the case of linear regression it can be hard to interpret the coefficients when one is of the scale 10^{-5} and others are much larger. When calculating the distance between points we also don't want it to be primarily determined by a single predictor because it has a much larger range. In the case of support vector machines the decision boundary should depend on the distribution of points and not their range. Similarly certain gradient based optimization methods take longer to converge when the "surface" is highly stretched out due to very different scales in the predictors.

For these reasons the data has been *centered* and *scaled* beforehand when the statistical method used demands it. By centering the data we transform it to have a mean of zero and scaling simply means dividing by the standard deviation.

$$x_{centered} = x - \bar{x} \tag{2.4}$$

$$x_{scaled} = \frac{x}{sd_x} \tag{2.5}$$

However the interpretation of coefficients can become much harder when applying these transformations. Centering affects the intercept but not the units. Whether one cares about this depends on the problem at hand. When scaling, the unit information is lost so we are suddenly no longer talking in degrees in the case of temperature. It does, however, make possible the comparison *between* coefficients.

2.2.3 Quirks and Anomalies

In the weather data, too, there are some anomalies. Notice that in table 2.3 above the maximum relative humidity is 101.04 percent! Luckily these anomalies do not occur often and the error is small. These values are left in without correction.

Highly correlated predictors

As one might suspect, many weather effects are highly correlated. More radiation generally means a higher value of sunshine. Similarly the three measures of temperature: minimum, maximum and average rise and fall in unison. Since methods like linear regression measure the proportional increase and all three are in the same units it can be safe to leave all but one out.

	avgTemp	maxTemp	radiation	sunshine	precipitation	humidity	windSpeed	windDirection	pressure
minTemp	**0.948**	0.870	0.445	0.175	0.156	-0.210	0.028	0.140	-0.132
avgTemp		**0.977**	0.630	0.378	0.078	-0.383	-0.113	0.089	-0.042
maxTemp			0.720	0.502	0.029	-0.474	-0.205	0.042	0.013
radiation				**0.867**	-0.177	-0.730	-0.244	-0.065	0.177
sunshine					-0.264	-0.706	-0.210	-0.144	0.266
precipitation						0.204	0.240	0.098	-0.340
humidity							-0.002	0.149	-0.132
windSpeed								0.123	-0.335
windDirection									-0.078

Table 2.4: Weather parameter correlations for full weather data-set

2.3 Combining weather and crime data

Some preliminary research was done to determine the viability of doing a more in depth investigation.

2.3.1 Distributions of high crime days

In order to see if days with particularly high crime tend to have specific weather conditions we divide the data into two groups: one consisting of 'normal' days and the other having all days where the number of crimes is equal or higher than one standard deviation above average.
For instance take the set of all days y which consists of the number of crimes of a particular category, say assault, on that day. With a total of n days we can say $y = \{y_1, y_2, y_i, \ldots, y_n\}$. This includes days where there is no crime, so where for some day i: $y_i = 0$. Now we map the data into a new set consisting of two categories; $\Delta = \{\delta_1, \delta_2, \delta_i, \ldots, \delta_n\}$ where $\delta_i \in \{Red, Green\}$ using the following rule:

$$\delta_i = \begin{cases} Red & \text{if } y_i \geq \bar{y} + sd_y \\ Green & \text{otherwise} \end{cases} \tag{2.6}$$

Figure 2.8 shows the results of applying this operation on the number of assaults in Amsterdam per day.[3] Taking the whole data-set gives uninteresting overlapping distributions. The data is further subdivided by season to see whether for instance sunshine matters more in the summer than it does in the winter. Note how especially for the average temperature (fourth row) the distributions are shifted towards warmer days for the high crime (red) subset. The effect is strongest in the spring and summer where the means are about 5 degrees apart and disappears in the winter and autumn.

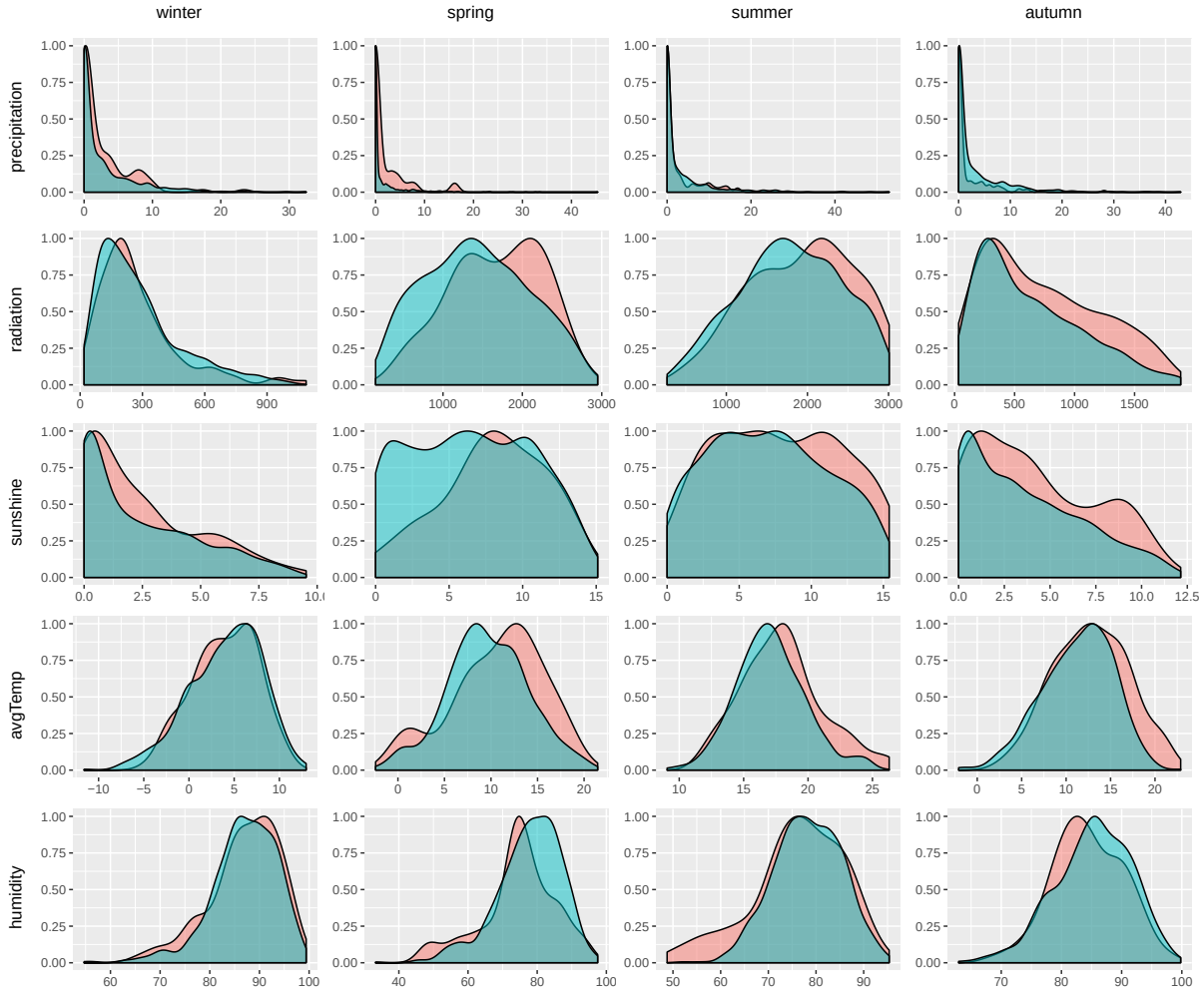


Figure 2.8: Amsterdam assault distributions per season and weather parameter

[3]Please see figure 8.3 in the appendix for the distributions for *robbery*, used later in the Amsterdam case study.

Chapter 3

Global Models

Global models try to approximate a function which can describe the entire feature space. This would be the most straightforward way to make predictions using new data.

3.1 Theoretical Descriptions

Before the results are shown we shall briefly describe the theoretical aspects of the algorithms used.

3.1.1 Linear Regression

The method of linear regression is a way of modeling the relationship between a dependent (response) value y and one or more explanatory values x with the addition of random noise, captured in ϵ. With the additional constant of $x_0 = 1$ we can write the expression more compactly as the product of two vectors:

$$\begin{aligned} y &\equiv \mathcal{F}(x) + \epsilon \\ \mathcal{F}(x) &= \beta^\intercal x \end{aligned} \tag{3.1}$$

The objective is to find some value of β (called $\hat{\beta}$) which minimizes the difference between the predicted values $\hat{y}$ and the actual values y. With this in mind, we define the "cost" of some estimate b as:

$$J(b) = \frac{1}{2m} \sum_{i=1}^{m} (b^\intercal x_i - y_i)^2 \tag{3.2}$$

where m is the number of observations in the data set. We can now write the minimization objective more succinctly as:

$$\begin{aligned} \hat{\beta} &= \arg\min_b J(b) \\ \hat{y} &= \hat{\beta}^\intercal x \end{aligned} \tag{3.3}$$

There are a variety of ways to find the minimum. There is an analytic solution called the *normal equation* but there also are iterative algorithms such as *gradient descent* (and its many variants) and *L-BFGS*. Usually a numerical solution is applied as it scales much better with the number of predictors: $O(kn^2)$ instead of $O(n^3)$ for the normal equation, where k is the number of iterations and n is the number of features [19].

3.1.2 Logistic Regression

It is possible to modify the equations of Linear Regression to get a classification model which describes the *probability* of an observation belonging to class 1:

$$\hat{y} = \hat{P}(y = 1 \mid x; \hat{\beta}) = 1 - \hat{P}(y = 0 \mid x; \hat{\beta}) \tag{3.4}$$

Where $y \in \{0, 1\}$. The algorithm needs to output a number in the range $(0 < \hat{y} < 1)$ instead of $\hat{y} \in \mathbb{R}$. For this we only need to wrap the former linear combination $\hat{\beta}^\intercal x$ in the *sigmoid function*:

$$\hat{y} = \frac{1}{1 + e^{-\hat{\beta}^\intercal x}} \tag{3.5}$$

we change the cost function to include both cases for y:

$$J(B) = -\frac{1}{m} \sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \tag{3.6}$$

The intuition here is that the cost goes to infinity when $\hat{y}$ approaches 1 while $y = 0$ and also when $\hat{y}$ approaches 0 while $y = 1$. In other words when the algorithm makes a very wrong prediction. Because the sigmoid function has a range of $(0, 1)$, the resulting cost is always finite, i.e. $\log(0)$ will never occur.

3.1.3 Count Models

Poisson Distributions

Because the data used in this project is essentially count data (the number of crimes on a certain day) it is natural to look for models which inherently have these assumptions built in. One such model is the Poisson model:

Poisson Regression Model

The Poisson model predicts the number of occurrences of an event. The dependent variable is a count (a non-negative integer). This makes sense in our case because you cannot have a negative number or a fraction of crimes on a day. For a Poisson distribution with a rate parameter μ, the probability that dependent variable Y will be equal to a certain number $y \in \mathbb{N}$ is:

$$p(Y = y) = \frac{e^{-\mu} \mu^y}{y!} \tag{3.7}$$

Where the Poisson regression assumption is that μ is dependent on the dot product between the parameters β and observations x:

$$\mu = e^{\beta^\intercal x} \tag{3.8}$$

The Poisson model has two important properties that are relevant here:

1. **Equidispersion property**: For the Poisson regression model it holds that the mean and variance are equal.

   $$E(y|x) = var(y|x) = \mu \tag{3.9}$$

   This property tends to be very restrictive and often fails to hold in practice. When the variance is greater than the mean it is called *overdispersion*, this is usually the case in real-life data. There is also *underdispersion*, where the counted outcome is mostly 0 or 1 [20]. In these instances it could be advisable to use a negative binomial model, explained below. Interestingly, the crime data shows both over- and underdispersion, depending on the crime category and municipality.

2. **Excess zero problem**: Usually the Poisson model will predict fewer zero's than are present in the data. It would be better to use a zero inflated Poisson or zero inflated negative binomial model when this situation arises.

Negative binomial distributions

In the case of over- or underdispersion of the data, a negative binomial model can be used. It is less restrictive than the Poisson model. An additional parameter α is introduced which characterizes the amount of overdispersion in the data.

$$var(y|x) = \mu + \alpha\mu^2 \tag{3.10}$$

When α is equal to zero we are left with the same equation as in the Poisson distribution above. Otherwise it is recommended to use the negative binomial model instead.

$$\begin{aligned} \alpha = 0 &\implies \text{Poisson model} \\ \alpha > 1 &\implies \text{overdispersion} \\ \alpha < 1 &\implies \text{underdispersion} \end{aligned}$$

Hurdle Models

In the case of excess zero's in the data one can specify a separate process that is responsible for generating the zero's and another one that will determine the count if the first process gives a "non-zero" answer. In other words, a Bernoulli probability distribution governs whether the outcome is zero or positive and a truncated at zero count model determines the distribution of the positives. This second count model could be, but is not limited to, a Poisson model or a negative binomial model.

When we define the process that generates the zero's as $f_1(.)$ and the process generating the non-zero positive answers as $f_2(.)$ the hurdle model is as follows:

$$g(y) = P(Y = y) = \begin{cases} f_1(0) & \text{if } y = 0 \\ \frac{1-f_1(0)}{1-f_2(0)} f_2(y) & \text{if } y \geq 1 \end{cases} \tag{3.11}$$

Note that the process $f_2(.)$ is *zero-truncated*, it still generates zero's but these are removed with the density function slightly modified to ensure that probabilities sum to unity [20]. An intuitive description is that first a "hurdle" must be overcome before the outcome is positive and a separate process is responsible for determining whether the hurdle is overcome or not.

Zero inflation

In the case of hurdle models we assumed that a separate process is responsible for determining zero or positive results. This may not be the case as the decision could come from two sources. Even though the first hurdle is overcome the result could still be zero for some other reason determined by the second process.

In some sense this distinguishes between "true" and "false" zero's. Where the true zero's are those generated by the failure to overcome the hurdle in the first place. The false zero's could in some sense be positive or belong to the non-zero event group.

As with the hurdle model the zero inflated model can be defined with the two processes $f_1(.)$ and $f_2(.)$:

$$g(y) = P(Y = y) = \begin{cases} f_1(0) + (1 - f_1(0)) f_2(0) & \text{if } y = 0 \\ (1 - f_1(0)) f_2(y) & \text{if } y \geq 1 \end{cases} \tag{3.12}$$

For this thesis, the *MASS* R package [21] was used for negative binomial models and the *PSCL* [22] package for the zero inflated models.

3.1.4 Support Vector Machines

Another approach of doing both classification and regression is using a Support Vector Machine (SVM). The essential idea in an SVM is the concept if *margin maximization*, where the decision boundary is placed in such a way that the distance between the points closest to the decision boundary is maximized. In figure 3.1 we can see an example of two decision boundaries; both separate the data but one provides a larger margin than the other [23].

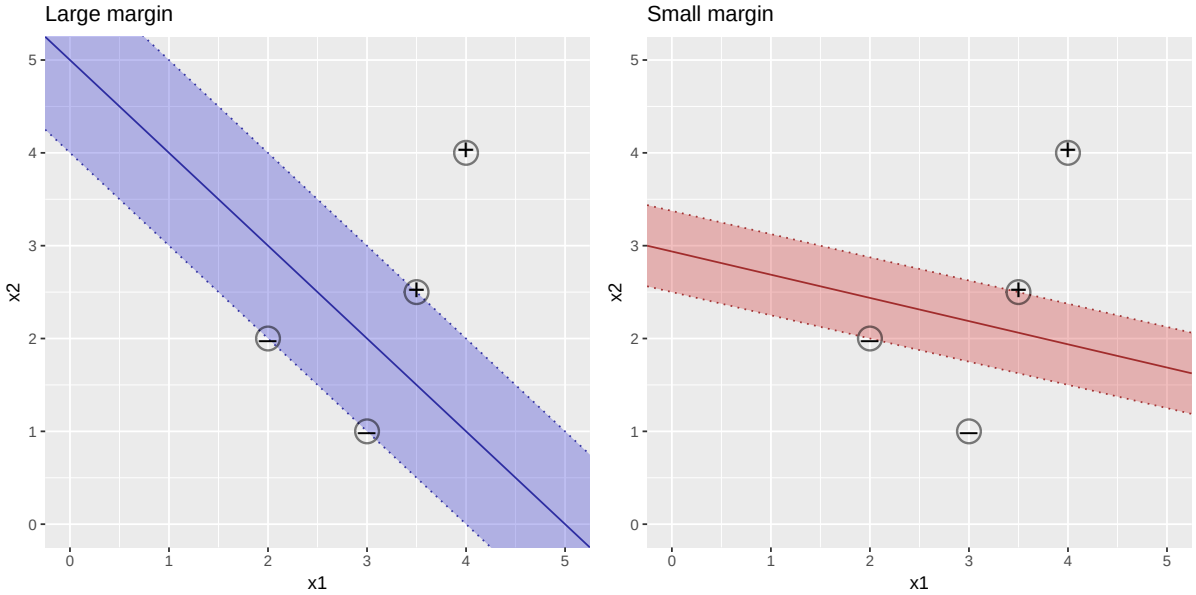


Figure 3.1: Support Vector Machine margins

Basic workings in binary classification

We would like to find the *decision boundary* that optimally separates the data-set. Begin by assuming that the data is linearly separable. Imagine a vector w that is perpendicular to the decision boundary and pointing in the direction of the "positive" points. We can write the decision boundary in terms of some vector w and scalar b:

$$w^\intercal x + b = 0 \tag{3.13}$$

Where $w^\intercal x + b < 0$ for "negative" points and $w^\intercal x + b > 0$ for "positive" points. To make it more convenient we introduce an extra variable y_n (the label of the point n):

$$y_n = \begin{cases} +1 & \text{for positive samples} \\ -1 & \text{for negative samples} \end{cases} \tag{3.14}$$

We can now write the constraint as a single equation. This is the same as saying that all points fall on the correct side of the dividing hyperplane. In other words, the data is linearly separable:

$$\forall n \ : \ y_n(w^\intercal x_n + b) \geq 0 \tag{3.15}$$

Lets define our notion of a margin. The margin $\hat{\gamma}$ with respect to a training example n can simply be written as:

$$\hat{\gamma}_n = y_n(w^\intercal x_n + b) \tag{3.16}$$

The intuition goes that, for a confident prediction, in the case of positive examples $w^\intercal x + b$ needs to be a large positive number and for negative examples it needs to be a large negative number. At this point there is a small problem: $\hat{y}$ is not a good measurement for confidence. Because y_n depends only on the sign, we can scale w and b and make $\hat{y}$ arbitrarily large without really changing anything meaningful. This implies that we need to put in some sort of normalization condition. There is more than one way to do this. Here we shall define the normalized margin γ with respect to a training example n to be of value 1 for the closest points [24]:

$$\gamma_n = \frac{y_n(w^\intercal x_n + b)}{\|w\|} \tag{3.17}$$

With the updated constraint:

$$\forall n \; : \; y_n(w^\intercal x_n + b) \geq 1 \tag{3.18}$$

The width of the margin is defined as $\frac{2}{\|w\|}$. Our optimization objective will be to maximize this quantity, which is the same as minimizing $\|w\|^2$:

$$\max \frac{2}{\|w\|} \implies \min \frac{1}{2}\|w\|^2 \tag{3.19}$$

Our new objective now is to find what w and b should be. We minimize the more convenient form $\frac{1}{2}w^2$ while respecting the constraints by using dual form and Lagrange multipliers:

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2}w^2 - \sum_{n=1}^{N} \alpha_n \Big[y_n(w^\intercal x_n + b) - 1 \Big] \tag{3.20}$$

To find the minimum we take the partial derivative with respect to w and set it to zero:

$$\nabla_w \mathcal{L} = w - \sum_{n=1}^{N} \alpha_n y_n x_n = 0 \tag{3.21}$$

Which implies that:

$$w = \sum_{n=1}^{N} \alpha_n y_n x_n \tag{3.22}$$

We also take the partial derivative with respect to b and set it to zero:

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{n=1}^{N} \alpha_n y_n = 0 \tag{3.23}$$

Which implies that:

$$\sum_{n=1}^{N} \alpha_n y_n = 0 \tag{3.24}$$

Now we can plug the results from the two partial derivatives back into $\mathcal{L}$ and simplify:

$$\mathcal{L} = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m x_n^\intercal x_m \tag{3.25}$$

The new optimization objective is to find the maximum of $\mathcal{L}$. This can be done with quadratic programming. Additionally, we figured out that the optimization depends only on the dot product of pairs of samples x_n and x_m. The quadratic programming software hands us back some vector α and it turns out most of the values (unless we're really unlucky) in α are zero. Where α is not zero, this corresponds to a *support vector*.

$$\alpha_n > 0 \implies x_n \text{ is a support vector} \tag{3.26}$$

In figure 3.1, these are the three left most points; the points that lie exactly on the margin boundary.

Finding w now depends solely on the support vectors SV. The size of SV is often a lot smaller than the total number of observations:

$$w = \sum_{x_n \in SV} \alpha_n y_n x_n \tag{3.27}$$

Determining b is also easy. Just take any support vector x_k, plug it in and solve for b:

$$y_k(w^\intercal x_k + b) = 1 \tag{3.28}$$

Up to this point the assumption has been that the data is perfectly linearly separable. This is obviously almost never the case. It is possible to allow some small errors by using a *soft margin*. Basically it gives the constraint some "slack" to move around. How much error is acceptable can be tweaked with a new parameter c.

Kernels

Another way to create a situation where the data is linearly separable is to apply some transformation ϕ to the data into a new space where it in fact *is* linearly separable and then use the mathematics described above in exactly the same way. This is possible because the optimization depends solely on combinations of pairs of samples x_n and x_m. It is not necessary to actually do the transformation, by replacing the dot product with another function (called a kernel) it is possible to implicitly do the transformation. This is called the *Kernel trick*.

Therefore it is not even necessary to "visit" this higher dimensional space. Interestingly it is possible to transform the data to a new space with an infinite number of dimensions, potentially without paying the costs of overfitting. Because in the end only a small number of support vectors define the decision boundary in the transformed high dimensional space.

Using no kernel at all is called the *linear kernel*. Some other popular kernels are:

Polynomial Kernel

Intuitively the polynomial kernel looks for combinations (sometimes called interactions) of features. A degree d polynomial is defined as:

$$K(x_n, x_m) = (x_n^\intercal x_m + c)^d \tag{3.29}$$

Radial Kernel

Another very popular kernel is the radial (or Gaussian) kernel. It can be seen as a similarity measure on the range $(0, 1)$ with value 1 when both points are the same and value 0 when they are infinitely far apart:

$$K(x_n, x_m) = \exp\left(-\frac{\|x_n - x_m\|^2}{2\sigma^2}\right) \tag{3.30}$$

A nice feature of the radial kernel is that it has an infinite number of dimensions since e can be written out as an infinite sum of terms. In the case of $\sigma = 1$:

$$K(x_n, x_m) = \exp\left(-\frac{1}{2}\|x_n - x_m\|^2\right) = \sum_{j=0}^{\infty} \frac{(x_n^\intercal x_m)^j}{j!} \exp\left(-\frac{1}{2}\|x_n\|^2\right)\exp\left(-\frac{1}{2}\|x_m\|^2\right) \tag{3.31}$$

The SVM's applied in this work all made use of the radial kernel.

3.1.5 Support Vector Machine Regression

It is possible to use a Support Vector Machine for regression by introducing a new parameter for tolerance called ϵ, where all residuals are smaller than or equal to ϵ. In figure 3.2 the blue band is 2ϵ wide. Points that fall outside the range are marked as x [25, 24].

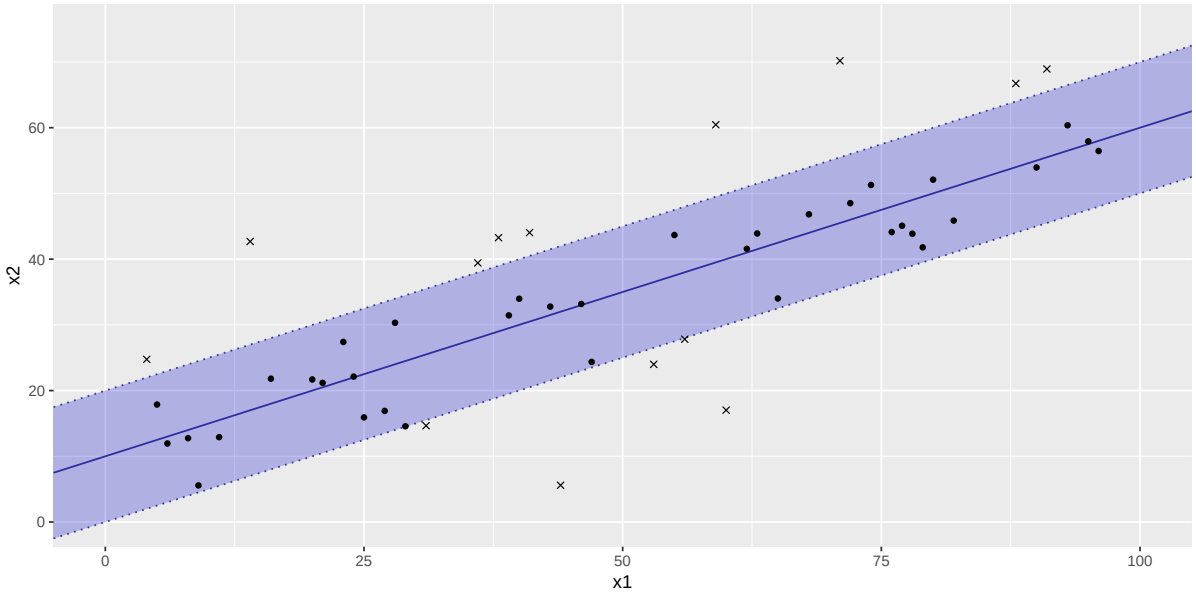


Figure 3.2: Support Vector Machine regression with ϵ margin

The same concept can be written as an extra set of constrains:

$$\forall n \; : \; |y_n - w^\intercal x_n + b| \leq \epsilon \tag{3.32}$$

Similarly as with the soft margin classifiers mentioned before, it is possible to add a slack variable to allow some points to fall outside the ϵ margin. In this project, the *e1071* R package [26] was used.

3.2 Setup of Experiments

3.2.1 Model Validation

Train-Test split

The data is split into two disjoint sets. One is used to train the model while the other is used to validate it. Due to the time series like character of the data and how future applications would use these models (use information from the past to make predictions about the future), it was decided that instead of a random sample, the data is split on January 1st 2012. This results in the years 2006 - 2011 being used for training and the years 2012 - 2013 being used for testing. The final three years (2014, 2015, 2016) were removed from the data set because they showed signs of being incomplete. All the global models make use of this particular split on January 1st 2012.

3.2.2 Performance Metrics

Of course we want to compare the model to some sort of baseline in order to see how well it's doing. Additionally we'd like a way of comparing different models with each other in a way that is independent of model complexity. We consider *Root Mean Squared Error* and *R-squared* when evaluating the regression models and we shall use *Precision*, *Recall*, *F-Score* and the *true positive/negative rates* to determine the performance of classification models.

Root Mean Squared Error

In order to compare between the different models we use a metric called *root mean squared error*. This number is simply a function of the difference between the predicted and actual values:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2} \tag{3.33}$$

This means it does not depend on the model complexity on the test set. However it is still in the units of the dependent variable and can therefore obfuscate the true error rate. One way of fixing this is to simply take the error as a percentage of the range of the dependent variable:

$$NRMSE = \frac{RMSE}{max(y) - min(y)} \tag{3.34}$$

We will include this *normalized* root mean squared error as the basis of comparing the efficacy of the different models discussed below.

R-squared

In the case of linear regression we also have access to *R-squared*. This number signifies the proportion of the variance in the dependent variable (in this case y) that is explained by the model. Usually this number lies on the interval (0, 1). It is always on this interval on the train sample, but in some cases it can be negative; for instance when the model predictions are worse than always predicting the average of y - denoted here as $\bar{y}$.

R-squared is the inverse of the ratio between the total variance (SST) and the residual sum of squares (SSR):

$$SSR = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 \tag{3.35}$$

$$SST = \sum_{i=1}^{m} (\bar{y} - y_i)^2 \tag{3.36}$$

$$R^2 = 1 - \frac{SSR}{SST} \tag{3.37}$$

Where m is the number of observations, $\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$ and $\hat{y}$ is the predicted value

Precision, Recall and F-Score

The classification models use precision, recall and F-Score instead of something more simple such as accuracy because a number like accuracy might be very deceptive. For instance, with a data set in which the vast majority of cases are class $\mathcal{A}$ and only very few belong to $\mathcal{B}$, simply predicting $\mathcal{A}$ regardless of any additional information would yield a very high accuracy.

- **Precision**
  This is the fraction of correctly classified observations of a certain class among all observations classified as such. So of all the points that the algorithm classified as class $\mathcal{A}$, what fraction actually was class $\mathcal{A}$?
- **Recall**
  The fraction of correctly classified observations of a certain class. In other words, of all the points belonging to class $\mathcal{A}$, what fraction did the algorithm classify as $\mathcal{A}$?
- **F-Score** It is sometimes convenient to combine precision and recall into a single number. This is what the F-Score does, it is the harmonic mean between the two:

$$F = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{3.38}$$

True positive/negative rates and Accuracy

The true positive and -negative rates are perhaps the most easy to understand: All they do is give us the proportion of correctly classified positives and negatives. In most cases (unless there is perfect separation of classes) there is a trade off between these two values. The true positive rate is in fact the same as recall described above. Accuracy is of course the fraction of correctly classified observations.

3.2.3 Baseline Models

As a comparison for the trained regression models we use two simpler models:

1. Always predicting the mean:
   As a simple baseline of comparison we can look at the root mean squared error of predicting a constant on the test sample, namely the mean of the dependent variable in the train sample. We would expect a model which is more complex than only predicting a constant (having more than just an intercept term) to do at least as well as this. Calculating the baseline is easy, with n observations in the train sample and m observations in the test sample:

$$\begin{aligned} RMSE_{mean} &= \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\bar{y}-y_i)^2} \\ \bar{y} &= \frac{1}{n}\sum_{j=1}^{n} y_j \end{aligned} \tag{3.39}$$

2. Predicting the value of a cyclical model:
   As we found in the data exploration phase, the number of crimes tends to fluctuate during the course of the year. For this reason we build a model that serves as a more realistic hypothesis: *Crime simply goes through an annual cycle and is independent of the weather*. This model lies between simply predicting the average and using a lot of complex weather related features. Firstly we calculate the period p of the dependent variable y by using the `spectrum` function in R. We would expect this to be 365.25 yet it turns out to be 364.5. This is probably due to missing days in the weather data set. Then we fit a linear regression model in R using the following formula:

$$\hat{y} = \sin\frac{2\pi}{364.5\,t} + \cos\frac{2\pi}{364.5\,t} \tag{3.40}$$

3.2.4 Transformations of the Dependent Variable

For some models doing a transformation on the dependent variable y could result in an increase in performance. Therefore the following basic transformations were attempted:

- **Exponential**
  Regression equation: $\log y = \beta^\intercal x$
  Predicted value: $\hat{y} = e^{\beta^\intercal x}$
- **Quadratic**
  Regression equation: $\sqrt{y} = \beta^\intercal x$
  Predicted value: $\hat{y} = (\beta^\intercal x)^2$
- **Reciprocal**
  Regression equation: $\frac{1}{y} = \beta^\intercal x$
  Predicted value: $\hat{y} = \frac{1}{\beta^\intercal x}$

3.2.5 Regression Data Structure

The structure of the data which has been used as the basis for predictions in Amsterdam is shown in table 3.1. With the explanatory variables (X) and dependent variable (y) given in solely numerical form.

X							y
Avg Temp	Humidity	Radiation	Precipitation	Pressure	Sunshine	Windspeed	Count
21.61430	82.33254	2101.5171	8.800305	1015.346	6.759527	3.475456	4
20.01682	74.45690	2245.8752	0.000000	1013.581	10.481544	5.847099	2
			⋮				⋮
17.41780	75.74595	1686.4503	0.000000	1016.878	7.328043	4.539718	0
16.38394	81.26223	1797.6301	6.653288	1008.539	8.188468	4.202404	6

Table 3.1: Weather and crime data used for regression

3.2.6 Classification Data Structure

In the case of classification the class we are trying to predict is whether the day is a high crime day or not. The definition of a high crime day is the same as in figure 2.8 with class = 1 if the number of crimes is one standard deviation above the mean.

X							y
Avg Temp	Humidity	Radiation	Precipitation	Pressure	Sunshine	Windspeed	Class
21.61430	82.33254	2101.5171	8.800305	1015.346	6.759527	3.475456	1
20.01682	74.45690	2245.8752	0.000000	1013.581	10.481544	5.847099	0
			⋮				⋮
17.41780	75.74595	1686.4503	0.000000	1016.878	7.328043	4.539718	0
16.38394	81.26223	1797.6301	6.653288	1008.539	8.188468	4.202404	1

Table 3.2: Weather and crime data used for classification

3.3 Results of Experiments

3.3.1 Regression

The following tables give the results of fitting a linear regression model to the data. Three different data-models are used:

- **Daily**: This is the most simple model, using only daily weather as shown in table 3.1
- **History**: Instead of the daily weather we take the mean of the three previous days as the predictor. This is more in line with the stress hypothesis where it takes a few days of uncomfortable weather to tip the balance and increase crime.
- **Seasonal**: By adding a season factor the model is trained using the weather variables per season. Because weather itself is highly seasonal it might have different effects depending on the time of year, e.g. relative coldness might be relaxing in the summer but debilitating in winter.

In table 3.3 we see the r-squared values of all the different data-models and transformations and, coincidentally, that they're not that great. The seasonal column gives us the best results at 4% explained variance.

	Daily	History	Seasonal
No Transformation	0.015	0.011	0.035
Exponential	0.017	0.013	0.046
Quadratic	0.018	0.015	0.048
Reciprocal	0.008	0.007	0.030

Table 3.3: R-squared for Linear Regression

When we look at the root mean squared errors the reciprocal transformation in combination with linear regression performs best with a reduction of about 16% over the baseline.

	avg	sin	glm	nb	zinb	svm
			Root Mean Squared Error			
No Transformation	4.003	4.078	4.107	4.317	4.101	3.808
Exponential			3.537			3.767
Quadratic			3.690			3.770
Reciprocal			**3.415**			3.793
			Normalized RMSE			
No Transformation	0.020	0.019	0.020	0.021	0.020	0.019
Exponential			0.017			0.018
Quadratic			0.018			0.018
Reciprocal			0.017			0.018
			Relative RMSE			
No Transformation	1	1.005	1.012	1.064	1.011	0.939
Exponential			0.872			0.927
Quadratic			0.910			0.928
Reciprocal			**0.842**			0.932

Table 3.4: Amsterdam Assault Daily

As shown in table 3.5, in the case of the history data-model there is a very slight but negligible increase in performance. The reciprocal transformation performs best in combination with the seasonal data-model. Interestingly in all cases linear regression has the minimum error and only in the case of no transformation the support vector machines do better.

Additionally the zero inflated model failed to converge in the seasonal case, perhaps due to a lack of zero cases or numerical instability in the data.

History data-model

	avg	sin	glm	nb	zinb	svm
			Root Mean Squared Error			
No Transformation	4.056	4.078	4.055	4.317	4.054	3.833
Exponential			3.528			3.808
Quadratic			3.673			3.810
Reciprocal			**3.411**			3.816
			Normalized RMSE			
No Transformation	0.020	0.019	0.019	0.022	0.019	0.018
Exponential			0.018			0.018
Quadratic			0.018			0.018
Reciprocal			**0.018**			0.018
			Relative RMSE			
No Transformation	1	1.005	1	1.064	0.999	0.945
Exponential			0.870			0.939
Quadratic			0.906			0.939
Reciprocal			**0.841**			0.941

Seasonal data-model

			Root Mean Squared Error			
No Transformation	4.003	4.078	4.220	4.585	*NA*	3.709
Exponential			3.613		*NA*	3.710
Quadratic			3.803		*NA*	3.717
Reciprocal			**3.456**		*NA*	3.685
			Normalized RMSE			
No Transformation	0.020	0.021	0.021	0.022	*NA*	0.018
Exponential			0.018		*NA*	0.018
Quadratic			0.019		*NA*	0.018
Reciprocal			**0.017**		*NA*	0.018
			Relative RMSE			
No Transformation	1	1.005	1.054	1.065	*NA*	0.914
Exponential			0.891		*NA*	0.915
Quadratic			0.938		*NA*	0.916
Reciprocal			**0.852**		*NA*	0.908

Table 3.5: Amsterdam Assault History and Seasonal

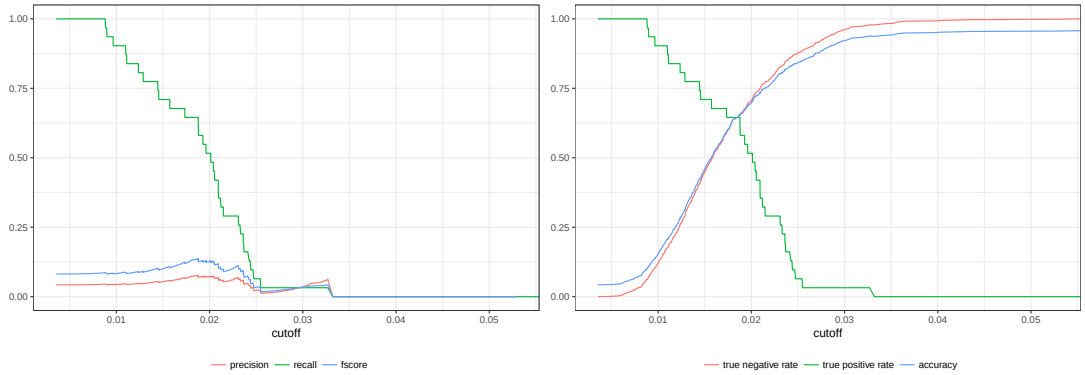
3.3.2 Classification

Below in figures 3.3a and 3.3b we can see the result of using the logistic regression algorithm on the data set. All the performance metrics are plotted as a function of *cutoff*: at which probability-value output we say an observation belongs to the class 1.

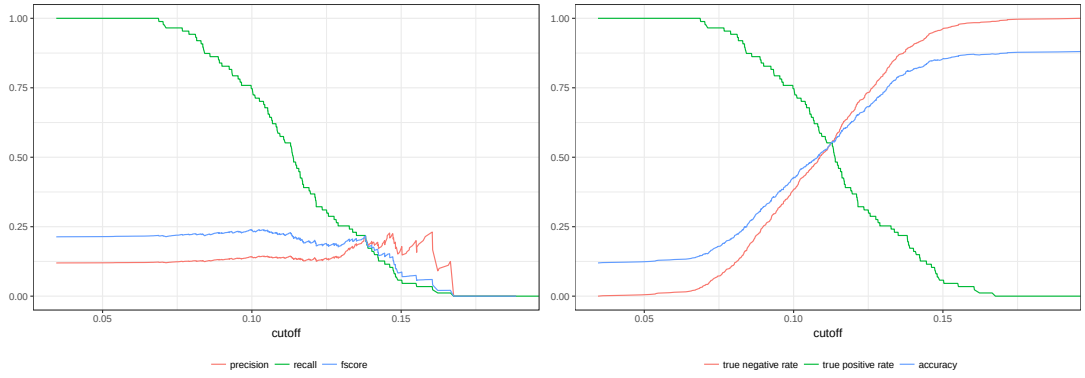
Note that in all cases the algorithm gives a very low probability - [0%, 3%) for assault and [0%, 16%) for robbery - to any observation belonging to class 1, i.e. being a high crime day. This also occurred when fitting a support vector machine to the same data. Even with a radial kernel and a lot of room to fit a complex decision-boundary the resulting fit always predicted class 0 on the test set.

Luckily the probability-like output of logistic regression provides us with some wiggle room to see how the algorithm performs on new data. There is as usual a trade-off between the true negative and true positive rates and where they intersect is also the point where f-score is maximized. This comes at a cost however, as we might capture most of the high crime days there are a lot of false positives.

For assault the maximum f-score is about 13% and for robbery it is a little less than 25%. This could indicate that robberies are easier to classify than assaults, which is in contradiction to what one would expect from the figures 2.8 and 8.3.



(a) Logistic regression output on assault data



(b) Logistic regression output on robbery data

Chapter 4

Local Models

Subgroup discovery is a well established data mining technique for finding descriptions of subsets of the data according some property of interest, formalized by a quality function, where interestingness is defined as *a distributional unusualness with respect to some property of interest* [27]. A common search strategy in subgroup discovery is to start with complete data and then iteratively add conditions and remove observations, using beam-search to find high quality subgroups.

4.1 Patient Rule Induction Method

The Patient Rule Induction Method (PRIM) is characterized by a very "patient" approach, i.e. it is not greedy and only removes a tiny bit of the data at each step while having a beam-width of 1. The property of interest (i.e. the *quality function*) is defined as the mean of some designated target variable. Note that in the general case, the quality function could be anything. PRIM can be applied for obtaining an overview of the relations in the data, for automatic hypothesis generation and for data exploration. Typically the resulting patterns are easy to interpret by users and domain experts [28]. Where each box $\mathcal{B}_k$ is a sub region consisting of logical conditions [29], for example:

$$\mathcal{B}_k = \begin{cases} 13.6 \leq \text{temperature} \leq 21.3 \ \& \\ \vdots \\ \text{precipitation} \leq 0.5 \end{cases} \tag{4.1}$$

In the case of PRIM, these algorithms go through a number of phases where the boxes are found, slightly edited and then validated. These phases are discussed below.

4.1.1 Top-down Peeling

Starting out with a box $\mathcal{B}$ that covers all of the data, at each iteration a small sub-box b within the current box $\mathcal{B}$ is removed. The number of observations which satisfy the box conditions is also called the *support*. This process is repeated until the support of $\mathcal{B}$ has reached some minimum quantity β_0. The sub-box b that is chosen is the one which maximizes the quality function.

For real valued inputs there are two possible sub-boxes, which take away the lower or upper portion of the subset. For some value of α and the jth input, the two eligible boxes are:

$$\begin{aligned} b_{j-} &= \{x \mid x_j < x_{j(\alpha)}\} \\ b_{j+} &= \{x \mid x_j > x_{j(1-\alpha)}\} \end{aligned} \tag{4.2}$$

Where $x_{j(\alpha)}$ is the α-quantile of the x_j-values in the current box. In the case of a categorical feature x_j there exists a candidate sub-box for each of its values s_{jm} within the current box:

$$b_{jm} = \{x \mid x_j = s_{jm}\} \tag{4.3}$$

Algorithm 1 gives the pseudo-code for the peeling process. With arguments:

1. minimum support β_0
2. quality function $\mathcal{Q}$ (the mean in the case of PRIM)
3. the data set $\mathcal{X}$
4. target variable y

Algorithm 1 Top-down Peeling

procedure PEEL(β_0, $\mathcal{Q}$, $\mathcal{X}$, y)
  $\mathcal{C} \leftarrow \emptyset$ ▷ empty set of conditions
  while $|\mathcal{X}| \geq \beta_0$ **do**
    $c \leftarrow C(\mathcal{X})$ ▷ calculate a set of candidate conditions for current $\mathcal{X}$
    $c^* \leftarrow \arg\max_c \mathcal{Q}(y)$ ▷ choose the condition that maximizes the quality function
    $\mathcal{X} \leftarrow \mathcal{X} \setminus \mathcal{X}[c^*]$ ▷ remove the observations that match the condition from $\mathcal{X}$
    $y \leftarrow y \setminus y[c^*]$ ▷ remove the observations that match the condition from y
    $C \leftarrow C \wedge c^*$ ▷ add the condition to $\mathcal{C}$
  return $\mathcal{C}$ as the optimal box

Note how a set of *conditions* $\mathcal{C}$ is returned instead of a set of (indexes of) observations b, as we are not particularly interested in the observations that fall into a box but more so the conditions that are required to construct it.

4.1.2 Bottom-up Pasting

The resulting box from the peeling process has been determined by the particular values that happened to define that sub-box at the various stages of the process. The decisions made in these stages were made without knowledge of later peels that further refined the the boundaries. Essentially the pasting process is the reverse of the peeling algorithm, a small sub-box b is added to the box $\mathcal{B}$. Again the sub-box b which maximizes the quality function is chosen, provided that quality increases.

4.1.3 Validation

After a box has been constructed it has to be checked against unseen data to make sure that the reported high quality is not the product of overfitting. To validate a box, we simply go through all the conditions that were constructed during the learning process and apply them one by one on a new data set. In figure 4.1 this process is visualized. The validated box consists of all the conditions up to and including the one that gives maximum quality.

4.1.4 Covering

In subgroup discovery it is customary to follow a so-called covering strategy. This means that the rule construction algorithm is applied sequentially to subsets of the data. First a rule is generated on the entire data set, afterwards all observations that fall into that box are removed. This process is repeated until the quality of the remaining subset becomes smaller than the overall quality. A pseudo code example is given in Algorithm 2.

For the covering procedure the arguments are:

1. minimum support β_0
2. quality function $\mathcal{Q}$ (the mean in the case of PRIM)
3. the data set $\mathcal{X}$
4. target variable y

Algorithm 2 Covering strategy

procedure COVER(β_0, $\mathcal{Q}$, $\mathcal{X}$, y)
  $q^* \leftarrow \mathcal{Q}(y)$ ▷ global quality
  $B \leftarrow \emptyset$ ▷ empty set of boxes
  while $\mathcal{Q}(y) \geq q^*$ **do**
    $b \leftarrow \texttt{peel}(\beta_0, \mathcal{Q}, \mathcal{X}, y)$ ▷ calculate the optimal box for current $\mathcal{X}$
    $B \leftarrow B \cup b$ ▷ report b as one of the covering boxes
    $\mathcal{X} \leftarrow \mathcal{X} \setminus b$ ▷ remove the observations that fall into box b from $\mathcal{X}$
    $y \leftarrow y \setminus b$ ▷ remove the observations that fall into box b from y
  return the covering boxes $\mathcal{B}$

In a slight abuse of notation, in algorithm 2 we let the `peel` procedure return references to specific observations. These can of course be trivially determined when the conditions defining a box are known.

4.2 Subgroup Discovery Package

Due to a lack of available transparent and well performing implementations of PRIM it was decided to code a new version in R. The most natural way of doing this is in the form of an R package, which also gives the possibility of releasing it on the *Comprehensive R Archive Network* (CRAN). The package was officially submitted and accepted by CRAN and can be used by anyone who wants to do subgroup discovery using the PRIM algorithm [30]. One advantage of this package is that it is written purely in R and has no dependencies on third party packages and is not simply a wrapper for existing libraries written in other languages.

There are a few differences with the pure PRIM description, namely the package has been written to be more generic and allow any quality function. Ties in quality while finding the best box are broken by the support of the boxes.

Additionally the package has many more features such as running parts in parallel, overcoming overfitting by giving a "2 standard errors below the optimum" option and constraining the greediness even more when dealing with logical and categorical inputs. Besides the covering strategy there is also a diversification algorithm, discussed below.

Currently the package does not include bottom-up pasting due to time constraints. The authors of PRIM also note that this strategy has a limited effectiveness in increasing the quality of a box [29]. Planned features are the bottom-up pasting algorithm and more subgroup discovery methods besides PRIM.

4.2.1 Training & Validation

Below in figure 4.1 we see an example of the training and validation code at work. Note that in this case the peeling process paints a very optimistic picture where the quality almost always goes up. The validation process shows us that we have in fact been overfitting and stopping earlier will yield a more generalized box. The horizontal dotted line shows the cutoff position when the "2 standard errors below the optimum" (2se) parameter is used. Of course the option to simply pick the highest quality box is also available.

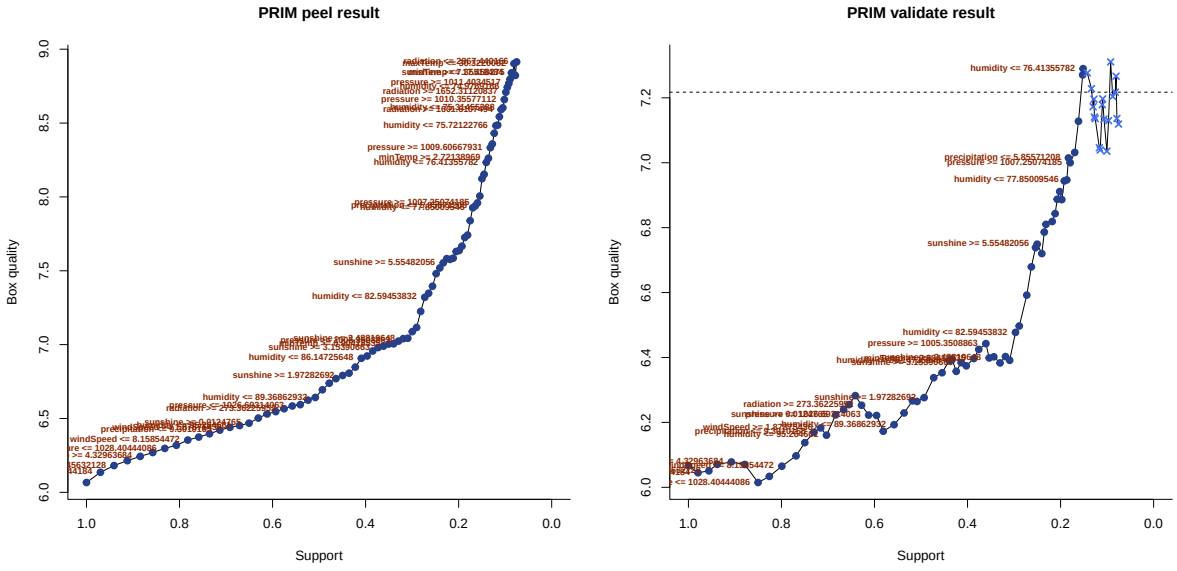


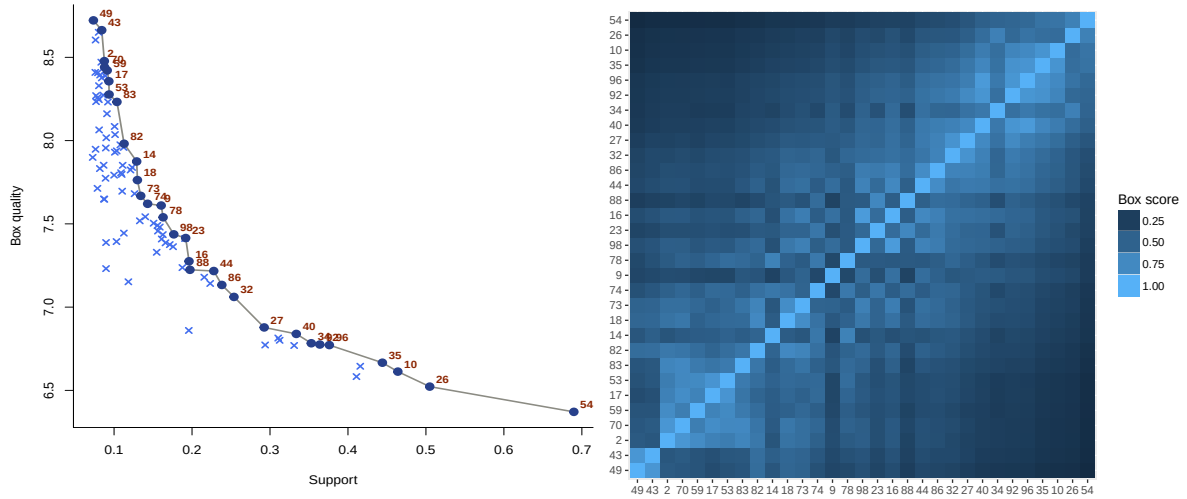
Figure 4.1: Training and validating the PRIM model

4.2.2 Diversification

In addition to the covering strategy, another possible approach of finding a diverse and interesting set of subgroups is to simply generate many independent random test-train splits of equal size and find the best box (according to whatever metric and cutoff point specified) on each one. This is a process that can easily be optimized by computing them in parallel. After all the boxes have been calculated, we only keep those that dominate the other boxes. In figure 4.2a we see all dominated boxes as crosses while the non-dominated boxes are labeled dots. (A point $p \in S$ is said to be non-dominated if there is no other point $q \in S$ whose coordinates are all greater than or equal to the corresponding coordinates of p.) To compare the "sameness" between boxes $\mathcal{A}$ and $\mathcal{B}$ we calculate a score by using the Jaccard index:

$$J(\mathcal{A}, \mathcal{B}) = \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \tag{4.4}$$

This equation causes the score to be close to 1 when the boxes both describe nearly the same subset and near 0 when almost no observation occurs in both $\mathcal{A}$ and $\mathcal{B}$. Interestingly, at least in case of the weather and crime data set, boxes that are close in quality and support also have a high similarity score. This is more clearly visible in figures 4.2a and 4.2b.



(a) Scatterplot highlighting non-dominated boxes

(b) Heatmap of Jaccard indexes

Figure 4.2: Comparing diversification sameness scores

Intuitively, if we suppose that points in figure 4.2a that are nearly equal in quality and support, also describe the same data, we would predict that in figure 4.2b the region around the diagonal line would be much brighter than the corners. As we can see this is indeed the case, in fact there is even some clustering going on. This could be an interesting direction for future work as these clusters describe different "ways" of being optimal. [1]

These results give us more confidence that the top scoring boxes have much in common and they are not simply sitting in their respective local maximums. Still, picking the best performing box is not always the best option. For instance in figure 4.2b, the two best performing boxes (lower left corner) sit in their own little cluster. We therefore prefer the third best box as it sits in a much larger cluster while still being very close to optimal and at the same time having a higher support.

[1] For a more clustered overview of figure 4.2b where the ordering has changed to put boxes with a high sameness score closer together please view figure 8.4 in the appendix.

4.3 Setup of Experiments

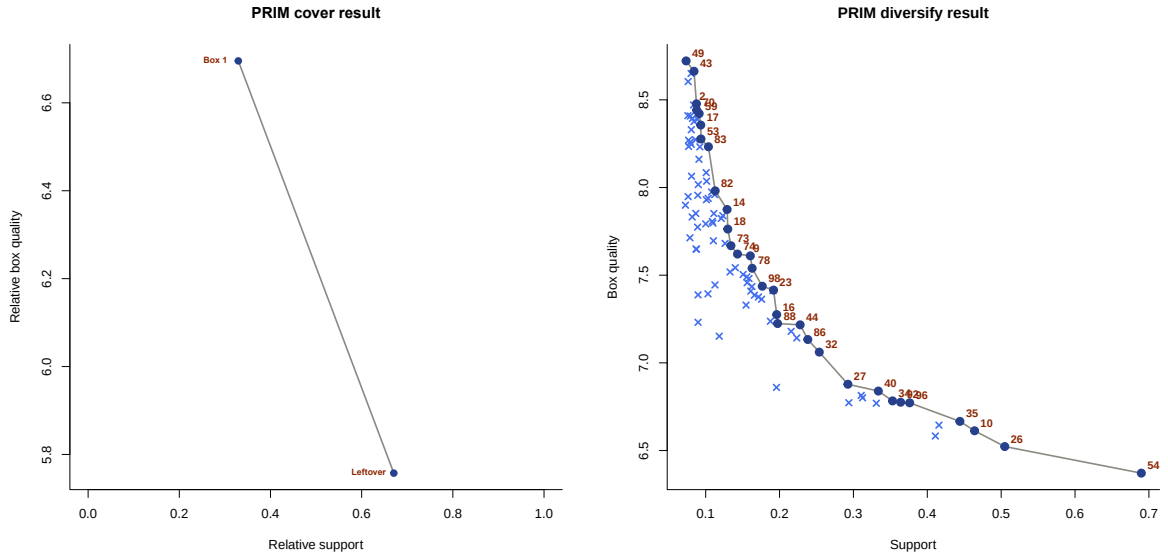
The data used for the local models is almost exactly the same as that for the global models: 10 years of daily weather for Amsterdam combined with daily sums of crimes. The only change is that the average temperature has been replaced by the minimum and maximum temperatures. This was done to make the final analysis easier to interpret.
Both the covering and diversification strategies were applied. To avoid overfitting, the optimal box was chosen using the "2se" rule and in the case of diversification, 100 different train-splits were attempted. All models have been validated using a test set that was not involved in training. These validated models where then applied on the whole data set. To further guard against overfitting, the final reported result is not the one with the best quality, but the optimal box in a well performing cluster, i.e. one with a high quality and (relatively) high support.

4.4 Results of Experiments

Both assault and robbery were analyzed for interesting local patterns and the results are described below. The subgroups found have a substantially larger average number of crimes per day while still describing around 275 days. For future work it will be trivial to include more crime categories due to the generic nature of the R package.

4.4.1 Assault

In figure 4.3 we can see the results of the covering (left) and diversification (right) strategies. The covering algorithm terminates quickly because the quality of the leftover observations after subtraction of the first box falls below the global average. We can think about this as removing a large fraction of the high crime days and being left with the normal days.
The diversification result yields many dominating boxes ranging from just above average but covering around 70% of the data to a very high 44% above average while only covering 7% of the data. Figure 4.2b has been constructed using the diversification data and can be consulted to find a good box. The first two best performing boxes have much in common, but they stand out from the others, we therefore pick the best box from the second cluster, box number two.



(a) Covering on the assault data set

(b) Diversifying on the assault data set

Figure 4.3: Results of applying PRIM to assault

Table 4.1 shows us the details of box nr 2 after validation from the diversification process. The quality is 1.4 times the global average and 287 days fall into the box (9% of the entire data set). As we would expect from a high performing box, the number of conditions is quite large and all 8 variables are used.

	relative	*absolute*
Box quality	1.4	8.48
Box support	0.09	287

Conditions				
		humidity	≤	75.29
8.73	≤	maxTemp		
3.19	≤	minTemp		
		precipitation	≤	0.11
1009.71	≤	pressure	≤	1027.97
1313.74	≤	radiation		
7.54	≤	sunshine		
1.96	≤	windspeed	≤	7.99

Table 4.1: Box nr 2 specifications

Looking at table 4.1 it is not straightforward how to interpret the result, i.e. what kind of day is described here? Therefore the box is visualized in figure 4.4 by taking the density function for each variable over the 10 years of observations in Amsterdam, where the red areas are those that fall into the box. We can see that in some cases, as with windspeed, a tiny slice at the left tail of the distribution is not covered. This shows us that this technique always requires someone to look at the result and decide whether this rule makes sense. In this case we might safely include the days with almost no wind, the bottom-up pasting process could help us here by enlarging the box and thereby removing some of the more spurious conditions.

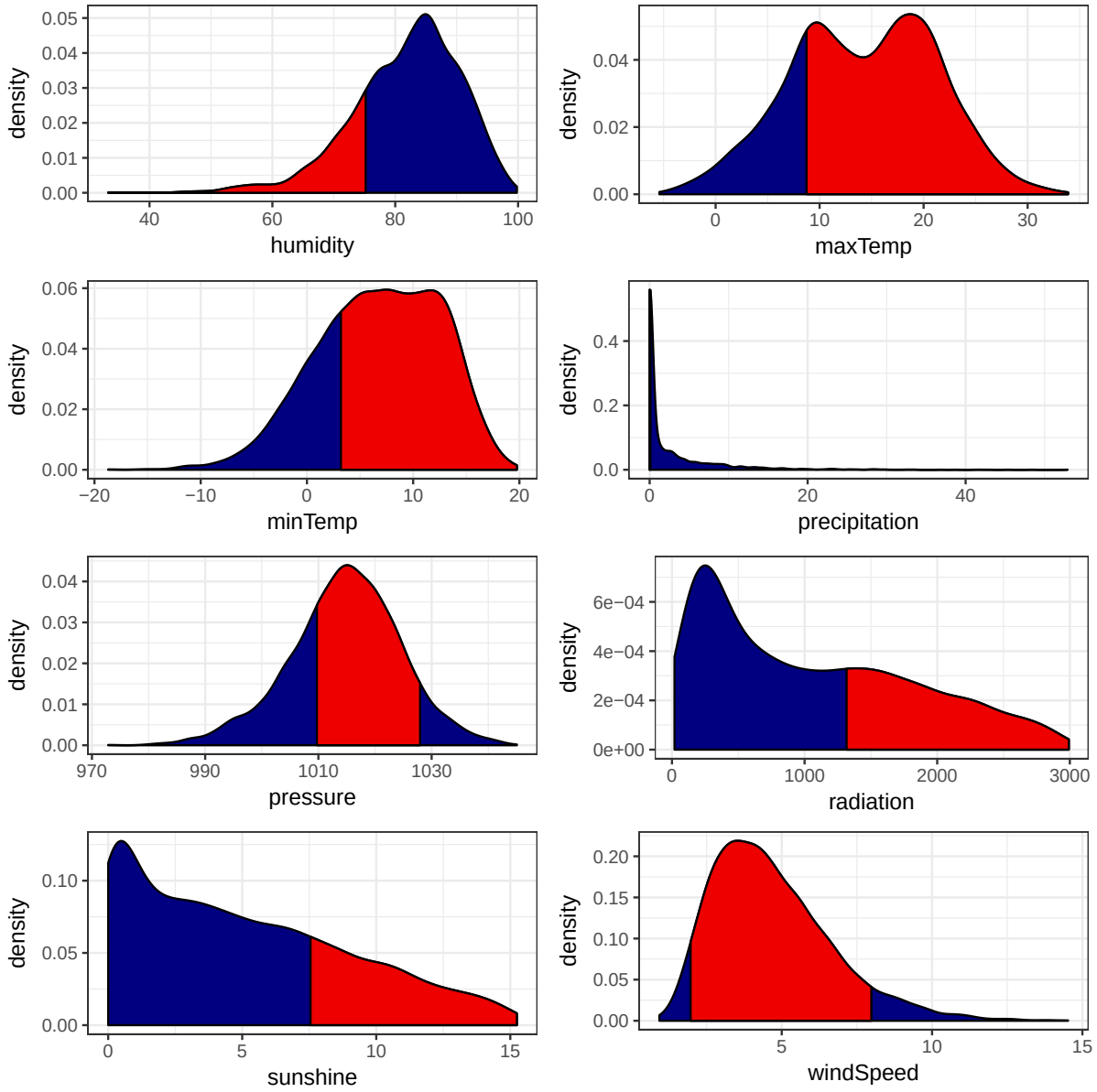


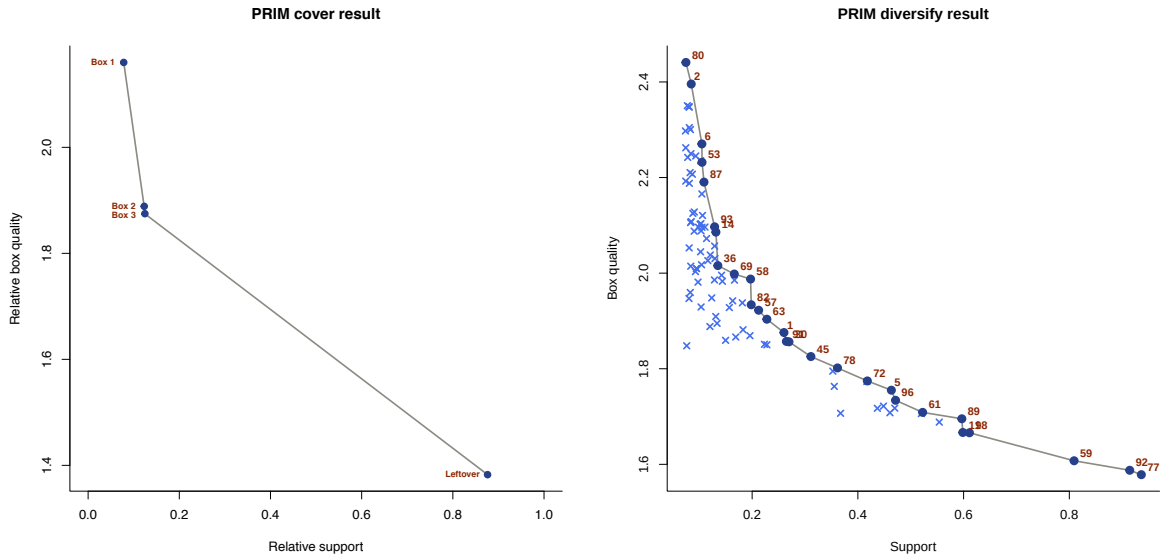
Figure 4.4: Rules of box nr 2 highlighted for assault

It is clearly visible that these distributions depict warm, dry and long days. This is in line with the related literature, which found a clear relation between temperature and violent crime [6].

4.4.2 Robbery

We do the same for robbery, this time box nr 6 is picked because boxes with the absolute best quality tend to introduce spurious conditions (e.g. removing a very small section from one tail of the distribution) which do not contribute to a better understanding of the subject material.

Covering yields a similar result with the first box having approximately the same quality as the best performing diversification boxes. Interestingly there are two more covers until the mean drops below the global average. This would suggest that the optimal box did not remove all the high quality observations.



(a) Covering on the robbery data set

(b) Diversifying on the robbery data set

Figure 4.5: Results of applying PRIM to robbery

In this case we can get an even better quality of 1.47 times the global average while covering 10% of the data, as is visible in table 4.2.

		relative		*absolute*
Box quality		1.47		2.27
Box support		0.1		344
		Conditions		
73.96	≤	humidity	≤	94.22
8.42	≤	maxTemp	≤	15.88
0.99	≤	minTemp	≤	10.08
		precipitation	≤	16.68
1003.26	≤	pressure	≤	1028.95
		radiation	≤	892.86
		sunshine	≤	6.19
2.64	≤	windspeed	≤	9.84

Table 4.2: Box nr 6 specifications

Again we can visualize this box by overlaying it on the variable distributions. Figure 4.6 shows us that in the case of robbery, a different pattern emerges. It are now dark and short days where the temperature never goes below freezing but it is not warm enough for a lot of people to go outside. We could speculate that this is due to a need for a minimum number of potential victims while constraining the maximum amount of witnesses. [2]

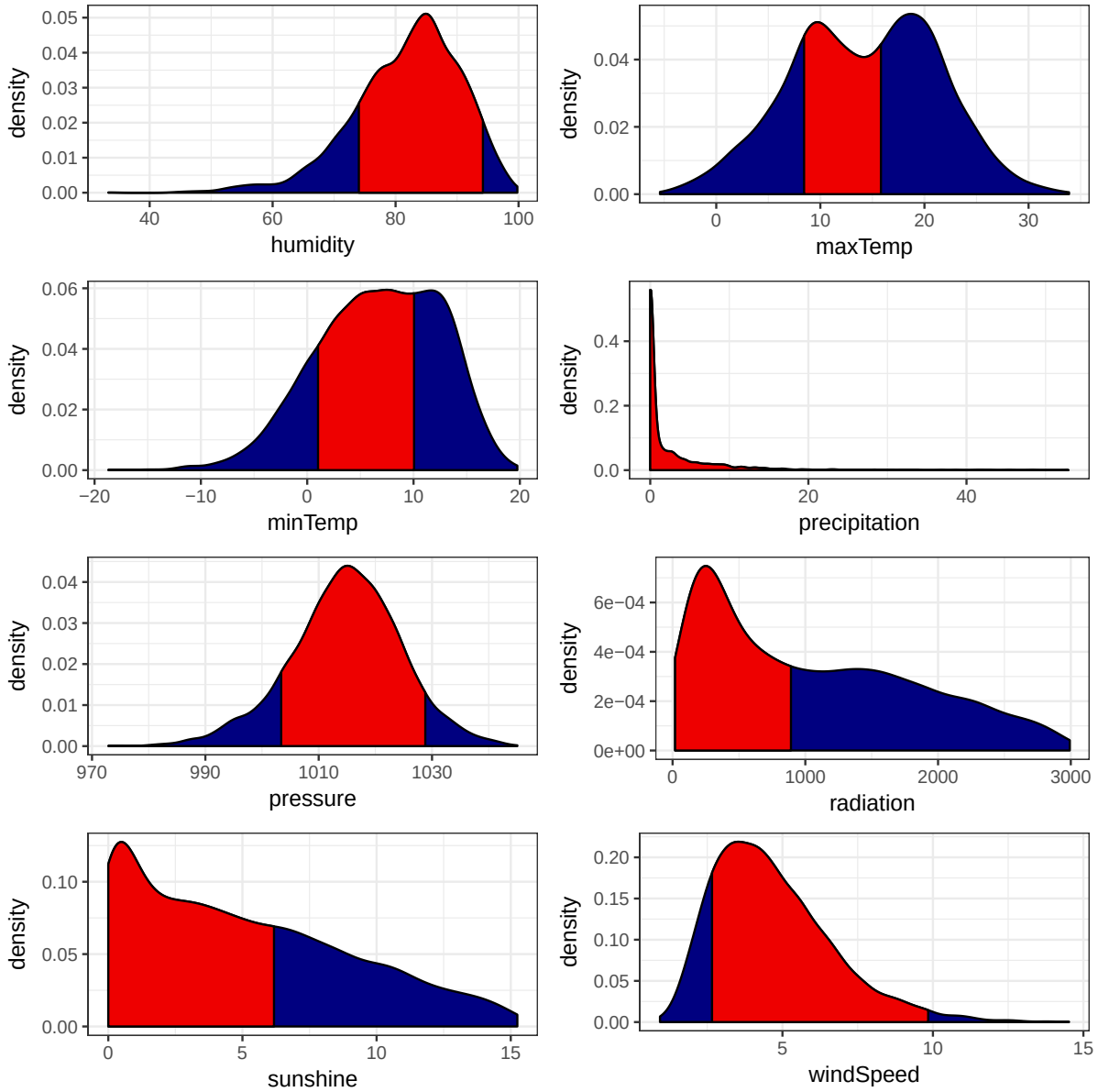


Figure 4.6: Rules of box nr 6 highlighted for robbery

[2]Please refer to table 8.1 in the appendix for an overview of which crimes are classified as robbery

4.5 Enhancing global models with local patterns

We can use the results from PRIM as an extra predictor in our global models. For this we shall first find two PRIM subgroups using the training set of the global models. One subgroup maximizes the average number of crimes and the other tries to minimize it. Then the new predictor becomes whether some observation falls into a box, for both of the PRIM boxes that were found.

Now we train the global models again but this time with the extra two variables (we shall call this the enhanced model). In table 4.3 we can see the coefficients and p-values of the log-linear (exponential) OLS model, trained on the assault data, both with and without the new explanatory variables. As expected for violent crime, temperature comes up as significant in both cases. In the enhanced model, precipitation almost becomes significant while both new PRIM variables are (highly) significant.

Looking at the estimated coefficients those with a very high p-value (low significance) often have the wrong sign, e.g. both radiation and sunshine should increase the number of assaults. However the more relevant and statistically significant predictors do have the correct sign: temperature increases the number of assaults and both PRIM predictors have their corresponding sign.

	Estimate	$Pr(>\|t\|)$
(Intercept)	$3.955e-01$	0.802
avgTemp	$7.482e-03$	0.016
radiation	$7.762e-06$	0.871
sunshine	$2.659e-03$	0.734
precipitation	$-4.407e-03$	0.135
humidity	$-2.274e-03$	0.338
pressure	$1.551e-03$	0.305
windSpeed	$-5.285e-04$	0.946

	Estimate	$Pr(>\|t\|)$
(Intercept)	$2.670e-00$	0.1114
avgTemp	$7.207e-03$	0.0198
radiation	$-1.957e-05$	0.6842
sunshine	$3.236e-03$	0.6776
precipitation	$-5.209e-03$	0.0778
humidity	$-1.112e-03$	0.6435
pressure	$-7.367e-04$	0.6478
windSpeed	$-2.138e-03$	0.7847
prim.maxTRUE	$1.305e-01$	0.0228
prim.minTRUE	$-2.008e-01$	$5.61e-05$

Table 4.3: Log-Linear Linear Regression coefficients

When we look at the changes in R-squared in table 4.4, a noticeable improvement is apparent. We can however still only account for a tiny fraction of the variance.

	Regular	Enhanced
No Transformation	0.015	0.025
Exponential	0.017	0.021
Quadratic	0.018	0.024
Reciprocal	0.008	0.010

Table 4.4: R-squared for Linear Regression on daily assaults

The RMSE performance is of the enhanced models is shown in table 4.5. Sadly there is a minimal change, which might indicate that there is a generalization problem. The SVM's do slightly better while the other models perform about the same.

	avg	sin	glm	nb	zinb	svm
			Root Mean Squared Error			
No Transformation	4.056	4.078	4.056	4.333	4.052	3.807
Exponential			3.541			3.767
Quadratic			3.683			3.770
Reciprocal			**3.433**			3.817
			Normalized RMSE			
No Transformation	0.020	0.019	0.020	0.022	0.020	0.019
Exponential			0.017			0.018
Quadratic			0.018			0.018
Reciprocal			**0.017**			0.019
			Relative RMSE			
No Transformation	1	1.005	1.000	1.068	0.999	0.939
Exponential			0.873			0.929
Quadratic			0.908			0.929
Reciprocal			**0.846**			0.941

Table 4.5: Amsterdam Assault Daily Enhanced

Chapter 5

Discussion

5.1 Global Models

It seems that the effects of weather on crime are very limited and difficult to describe using a global model. The amount of noise compared to the signal is very large. Partly this is caused by how the crimes were processed and reported by the police and the ministry of security and justice. The other part of the noise is more natural and mostly due to unpredictable human behavior. Opportunities for potential crimes are distributed randomly.

Several methods of pattern recognition have been applied where linear regression gives at best a 10% decrease in error over the baseline model. Only a small fraction of the variance could be explained, namely 1-4%. Other methods such as support vector machines and zero inflated negative binomial models did not do better.

Most likely important predictors are missing which could better explain the variance. Finding these explanatory variables could be a difficult task in itself. Human beings are influenced by many factors with people reacting differently to the same stimulus. At this point being able to accurately predict Kingsday is a much more effective way of curbing crime.

5.2 Local Models

The performance of the patient rule induction method seems to exceed those of any global model. This could be because a local model does not need to optimize for the entire feature space and can instead focus on one specific subset. The subsets found with the highest quality consistently boasted a 40-50% increase over the global quality while still containing a substantial number of days.

The results are in line with previous research, the subset we found with many violent crimes consists mainly of warm, summer days. Meanwhile the subset for economically motivated crimes (such as robbery) has many typical Dutch late autumn days.

Of course the question still remains how directly the weather can really influence crime. More likely the weather influences people's behavior which in turn influences crime. Additionally certain cultural phenomena occur during specific times of the year and could imply a false relationship between particular weather conditions and crimes, i.e. they are confounding variables.

Chapter 6

Future Work

6.1 Peak Matching

Instead of predicting the actual number of crimes per day or per month, more accuracy could be gained from predicting when crime goes up or down. Essentially turning it into a classification problem. In this case, algorithms like support vector machines seem to be bad at predicting the correct absolute number but are good at following the general peaks and troughs. However it is questionable how much information could be gained from this method. It would not be known by *how much* crime will rise, only that it will.

Another more complex variation of above idea is to predict the derivative or slope on a point in time. This would also tell us by how much crime will rise or fall, but still not an absolute amount because the starting conditions are unknown.

6.2 Extreme Weather Conditions

Storms and heatwaves could contribute to a real change in the number of daily crimes. It could be interesting to tell which effects they have on specific types of crime. For instance during storms most people stay indoors. During a heatwave perhaps many people open their windows which allows for easy access for burglars. Of course at this point this is just speculation.

6.3 Clustering PRIM Results

The PRIM diversify output lends itself well to a clustering analysis. Then one box from each cluster could be chosen and analyzed. This would yield more information on the different ways of being a high quality subset. For instance some of these clusters could represent days with extreme conditions mentioned before. Figure 8.4 in the appendix shows us a start on a method of extracting the clustering information.

6.4 Generalization of Subgroup Discovery Package

The current version of subgroup discovery is a limited generalization of the PRIM method where different quality functions can be chosen other than the mean. Further abstraction can yield a more general form of beam search subgroup discovery where PRIM is simply running the program with beam width set to 1 and quality function set to the mean.

6.5 Additional Data

6.5.1 Police Crime Data

The data used for this project was collected by the Dutch ministry of Security and Justice and only includes cases that have been through the entire court process, i.e. a judge has ruled on a verdict. This filtering limits the number of data points available and can obfuscate the true underlying processes. Are we for instance only capturing and convicting a certain type of criminal? An analogy to this problem would be interviewing people on the bus or train about their experiences on public transport; these answers cannot give you the full picture as you only gather data on subjects who made use of it in the first place.

6.5.2 Higher Resolution Data & More Complex Features

Including the time and postcode of a crime can yield more possibilities other than simply being more accurate on the weather conditions. We could for instance think of a *heat stress index*, a new feature constructed out of variables like temperature, humidity and pressure on a specific time.

6.5.3 Non-Weather Data

Data from other sources could be used in combination with weather data. Given a better resolution of when and where a crime took place this could be matched with facts about the general area such as income and wealth distribution and population density or even the number of trees and greenery.

Chapter 7

Conclusions

The goal of this research was to investigate how the number of crimes in a crime category depends on certain weather variables. Firstly by determining if a statistical relationship between weather and crime can be ascertained and secondly by building and testing predictive models.

The results gained from the global models show a definite but very weak signal which is hard to distinguish from the noise present in the data. At most a 16% increase in predictive performance over simple baseline models can be achieved when working with the assault crime category. Generalizing to new data was difficult for these models.

On the other hand, local models faired much better at finding high quality subgroups with characteristics that are in line with research elsewhere. A 50% increase in the average number of daily crimes was reached in the best performing subgroups while still describing a relevant fraction of the available days (±10%).

The results from these local models could be used to better inform law enforcement organizations about at what times of the year the number of crimes in a certain crime category are maximized.

Bibliography

[1] J. Horrocks and A. K. Menclova, "The effects of weather on crime," *New Zealand Economic Papers*, vol. 45, no. 3, pp. 231–254, 2011.

[2] L. E. Cohen and M. Felson, "Social change and crime rate trends: A routine activity approach," *American sociological review*, pp. 588–608, 1979.

[3] R. Wirth and J. Hipp, "Crisp-dm: Towards a standard process model for data mining," in *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, 2000, pp. 29–39.

[4] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017. [Online]. Available: https://www.R-project.org/

[5] C. A. Anderson and D. C. Anderson, "Ambient temperature and violent crime: Tests of the linear and curvilinear hypotheses." *Journal of personality and social psychology*, vol. 46, no. 1, p. 91, 1984.

[6] J. L. Cotton, "Ambient temperature and violent crime," *Journal of Applied Social Psychology*, vol. 16, no. 9, pp. 786–801, 1986.

[7] L. Ross and C. A. Anderson, "Shortcomings in the attribution process: On the origins and maintenance of erroneous social assessments," 1982.

[8] E. G. Cohn, "Weather and crime," *The British Journal of Criminology*, vol. 30, no. 1, pp. 51–64, 1990.

[9] H. Feldman and R. Jarmon, "Factors influencing criminal behavior in newark: A local study in forensic psychiatry," *Journal of Forensic Science*, vol. 24, no. 1, pp. 234–239, 1979.

[10] K. D. Harries and S. J. Stadler, "Determinism revisited: Assault and heat stress in dallas, 1980," *Environment and Behavior*, vol. 15, no. 2, pp. 235–256, 1983.

[11] K. D. Harries and S. Stadler, "Heat and violence: New findings from dallas field data, 1980–1981," *Journal of Applied Social Psychology*, vol. 18, no. 2, pp. 129–138, 1988.

[12] J. Rotton and J. Frey, "Air pollution, weather, and violent crimes: concomitant time-series analysis of archival data." *Journal of personality and social psychology*, vol. 49, no. 5, p. 1207, 1985.

[13] J. D. Perry and M. E. Simpson, "Violent crimes in a city: Environmental determinants," *Environment and Behavior*, vol. 19, no. 1, pp. 77–90, 1987.

[14] E. Baumer and R. Wright, "Crime seasonality and serious scholarship: A comment on farrell and pease," *Brit. J. Criminology*, vol. 36, p. 579, 1996.

[15] J. L. Lauritsen and N. White, *Seasonal Patterns in Criminal Victimization Trends*. US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics, 2014.

[16] D. W. Osgood, "Poisson-based regression analysis of aggregate crime rates," *Journal of quantitative criminology*, vol. 16, no. 1, pp. 21–43, 2000.

[17] M. Ranson, "Crime, weather, and climate change," *Journal of environmental economics and management*, vol. 67, no. 3, pp. 274–302, 2014.

[18] R. Sluiter, "Interpolation methods for the climate atlas," *KNMI technical rapport TR–335, Royal Netherlands Meteorological Institute, De Bilt*, pp. 1–71, 2012.

[19] J. Fox, *Applied regression analysis and generalized linear models*. Sage Publications, 2015.

[20] A. C. Cameron and P. K. Trivedi, *Microeconometrics: methods and applications*. Cambridge university press, 2005.

[21] W. N. Venables and B. D. Ripley, *Modern Applied Statistics with S*, 4th ed. New York: Springer, 2002, iSBN 0-387-95457-0. [Online]. Available: http://www.stats.ox.ac.uk/pub/MASS4

[22] A. Zeileis, C. Kleiber, and S. Jackman, "Regression models for count data in R," *Journal of Statistical Software*, vol. 27, no. 8, 2008. [Online]. Available: http://www.jstatsoft.org/v27/i08/

[23] Y. Abu-Mostafa. (2012) Learning from data. [Online]. Available: https://work.caltech.edu/telecourse.html

[24] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[25] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[26] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2017, r package version 1.6-8. [Online]. Available: https://CRAN.R-project.org/package=e1071

[27] S. Wrobel, "An algorithm for multi-relational discovery of subgroups," *Principles of Data Mining and Knowledge Discovery*, pp. 78–87, 1997.

[28] M. Atzmueller, "Subgroup discovery," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 5, no. 1, pp. 35–49, 2015.

[29] J. H. Friedman and N. I. Fisher, "Bump hunting in high-dimensional data," *Statistics and Computing*, vol. 9, no. 2, pp. 123–143, 1999.

[30] J. Baas, *subgroup.discovery: Subgroup discovery and bump hunting*, 2017, r package version 0.2.0. [Online]. Available: https://cran.r-project.org/web/packages/subgroup.discovery

Chapter 8

Appendix

Assault
Mishandeling
Mishandeling - algemeen
Mishandeling / GSB
Mishandeling - horecagew. arrond. Almelo
Mishandeling - willekeurig geweld
Overige mishandeling
Openlijke geweldpleging
Openlijke geweldpleging - algemeen
Openlijke geweldpleging / GSB
Openlijke geweldpleging - horecagew. arrond. Almelo
Openlijke geweldpleging - willekeurig geweld
Overige openlijke geweldpleging
Geweld tegen ambtenaren
Geweld tegen beroepsbeoefenaar ambulance
Geweld tegen beroepsbeoefenaars ambulance
Geweld tegen beroepsbeoefenaars brandweer
Geweld tegen beroepsbeoefenaars OV
Geweld tegen beroepsbeoefenaars overig
Geweld tegen beroepsbeoefenaars politie
Geweld tegen beroepsbeoefenaars ziekenhuizen
Geweld tegen een politie-ambtenaar
Geweld tegen medew. Airport Security/Douane/KMAR
Geweld tegen Politie
Zware mishandeling

Vehicle Theft
Autodiefstal
Diefstal brom-/snorfiets
Diefstal brom-/ snorfiets
Diefstal overige vervoermiddelen
Diefstal overige vervoermiddelen / GSB
Diefstal van auto
Diefstal van fiets
Diefstal van fiets / GSB
Diefstal van motor
Overige motorvoertuigendiefstal
Vaartuigdiefstal
Werkvoertuigdiefstal

Murder
Overige moord en doodslag
Moord en doodslag
Moord en doodslag - algemeen
Moord en doodslag - horecagew. arrond. Almelo
Moord en doodslag - willekeurig geweld

Rape
Verkrachting
Aanranding
Overig sexueel
Zedendelicten
Zedendelicten Algemeen
Zedendelicten / GSB
Zedendelicten - willekeurig geweld
Zedenzaak overig

Domestic Violence
Huiselijk geweld
Huiselijk geweld kindermishandeling
Huiselijk geweld oudermishandeling
Huiselijk geweld overig
Huiselijk geweld partnermishandeling

Robbery
Straatroof
Straatroof (waaronder tasjesroof)
Diefstal met geweld

Hard Drug
Harddrugs
Harddrugs / GSB
Harddrugs overig handel en smokkel
Harddrugs overig productie
Overige harddrugs
Synthetische drugs
Cocaine handel smokkel
Heroine handel smokkel
Herone/Cocane
Straathandel / drugsrunners
Overige drugsdelicten”

General
Overige
Overige wetten

Burglary
Inbraak in bedrijf / kantoor
Inbraak in school
Inbraak in woning
Inbraak overige objecten
Woninginbraak
Insluiping woning

Larceny
Diefstal
Diefstal af / uit overige voertuigen
Diefstal uit bedrijf
Diefstal uit overige objecten
Diefstal uit / vanaf auto
Diefstal uit woning
Brandstofdiefstal
Eenvoudige diefstal
Kentekenplaatdiefstal
Koperdiefstal
Ladingdiefstal
Overige diefstal
Overige diefstallen
Overige diefstallen / GSB
Overig eenvoudige diefstal
Overig gekwalificeerd diefstal
Winkeldiefstal

Soft Drug
Softdrugs
Softdrugs / GSB
Hash handel en smokkel
Hash productie
Coffeeshop gerelateerd
Handelingen i.h.k.v. exploitatie coffeeshop
Handelingen i.h.k.v. exploitatie coffeeshop

Discriminatie
Discriminatie antisemitisme
Discriminatie godsdienst / levensovertuiging
Discriminatie godsdienst/ levensovertuiging
Discriminatie handicap
Discriminatie overig
Discriminatie ras
Discriminatie seksuele gerichtheid

Table 8.1: Crime super-category allocation

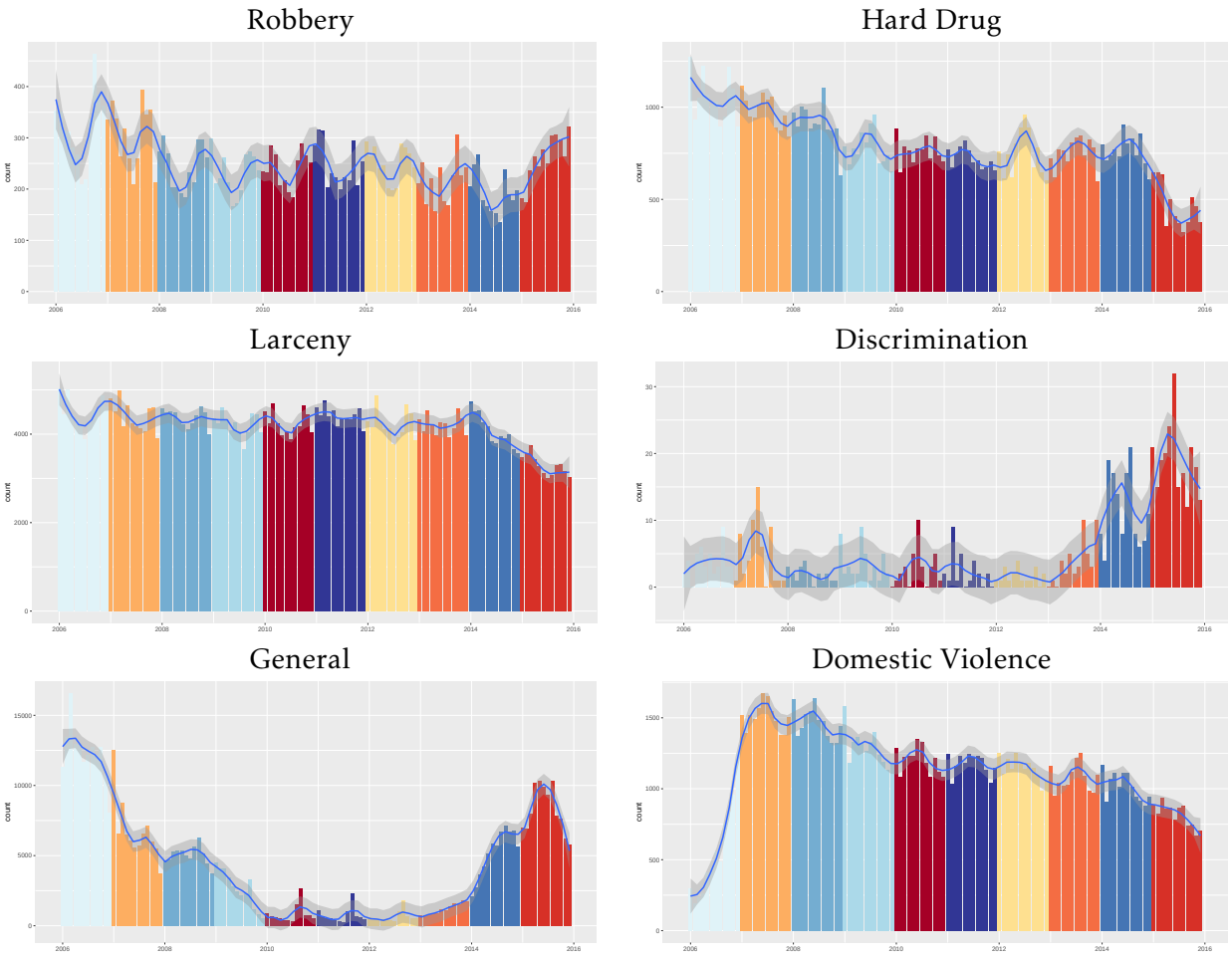


Figure 8.1: Monthly crime counts per category

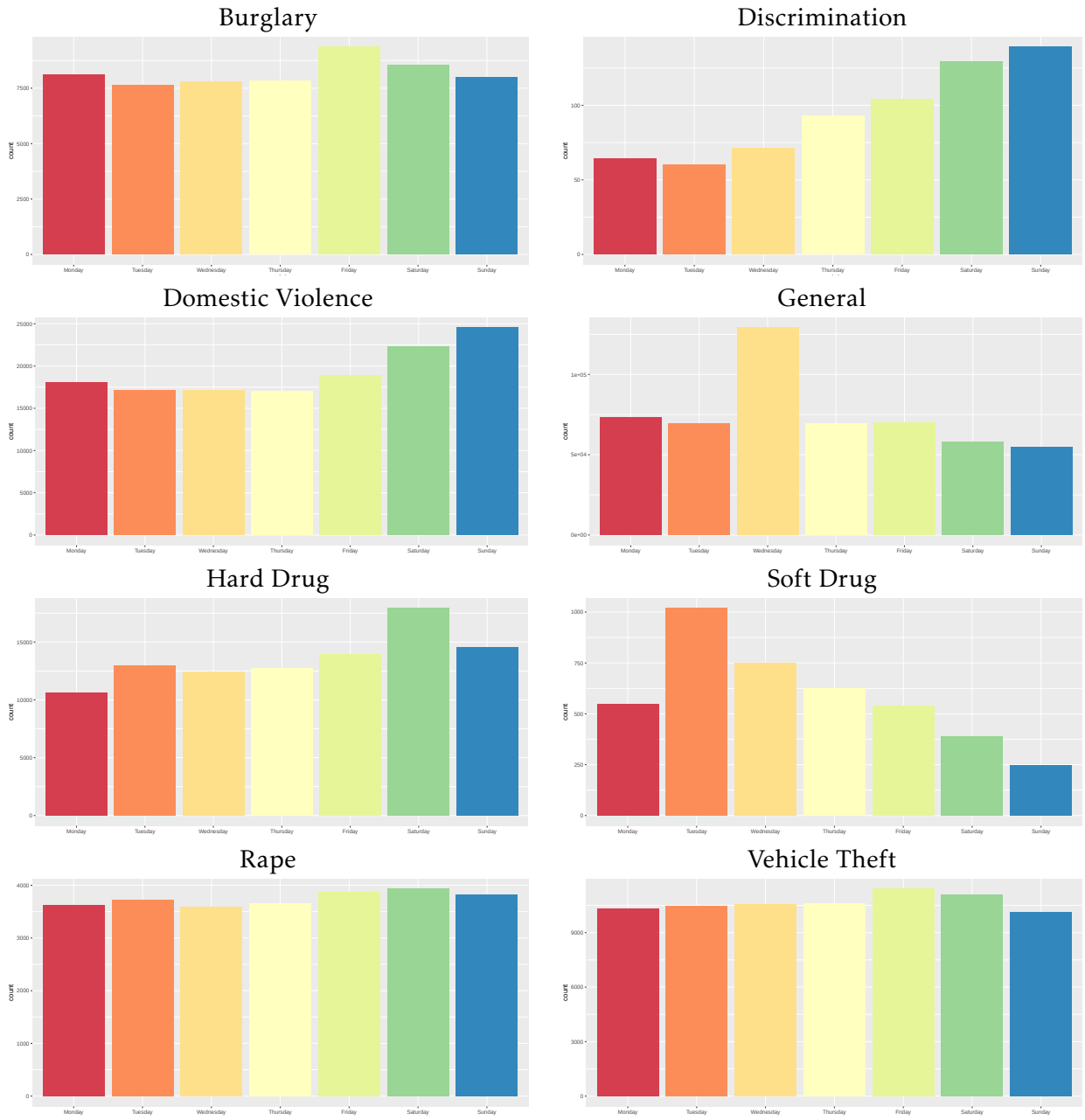


Figure 8.2: Weekday crime distributions per category

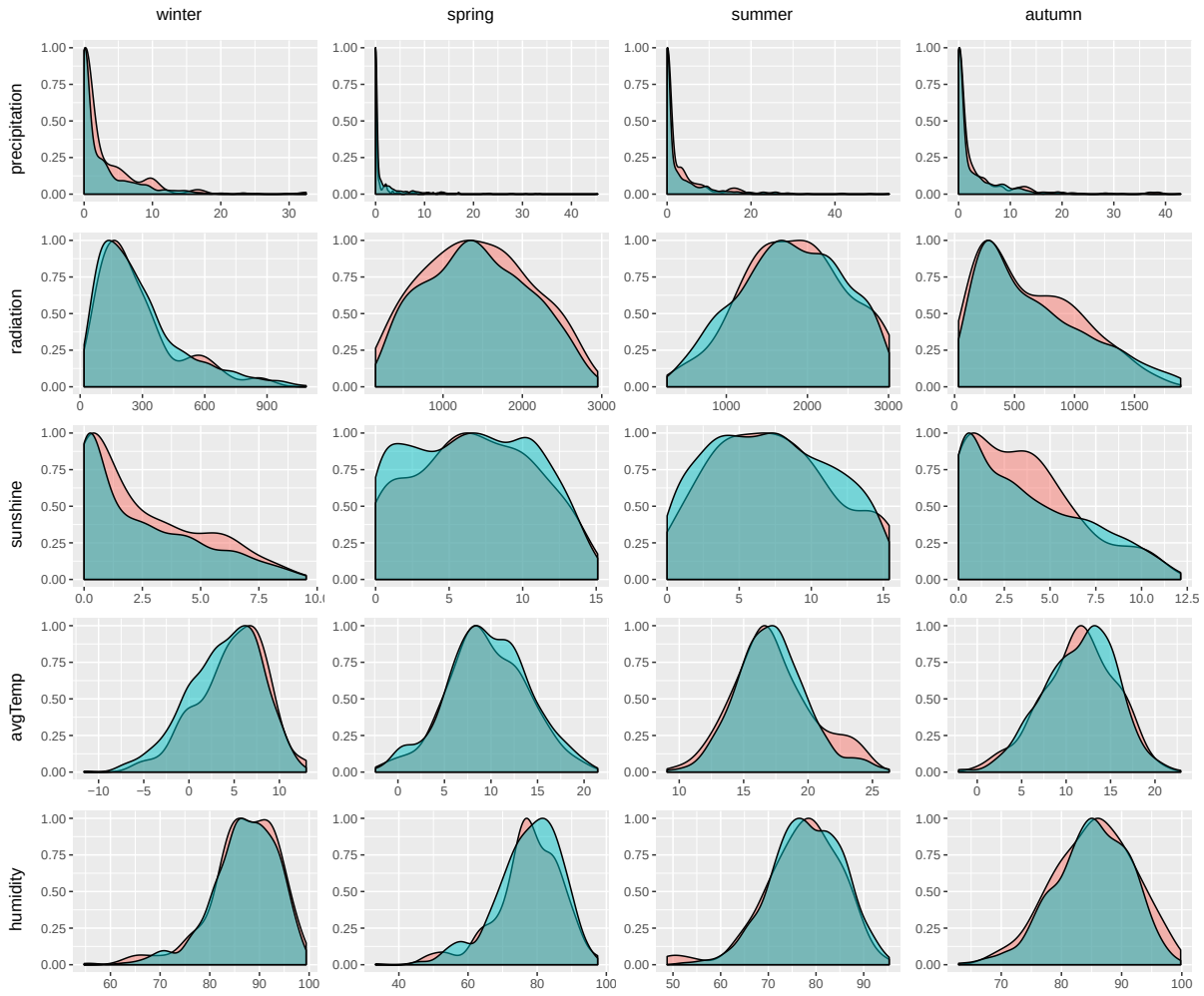


Figure 8.3: Amsterdam robbery distributions per season and weather parameter

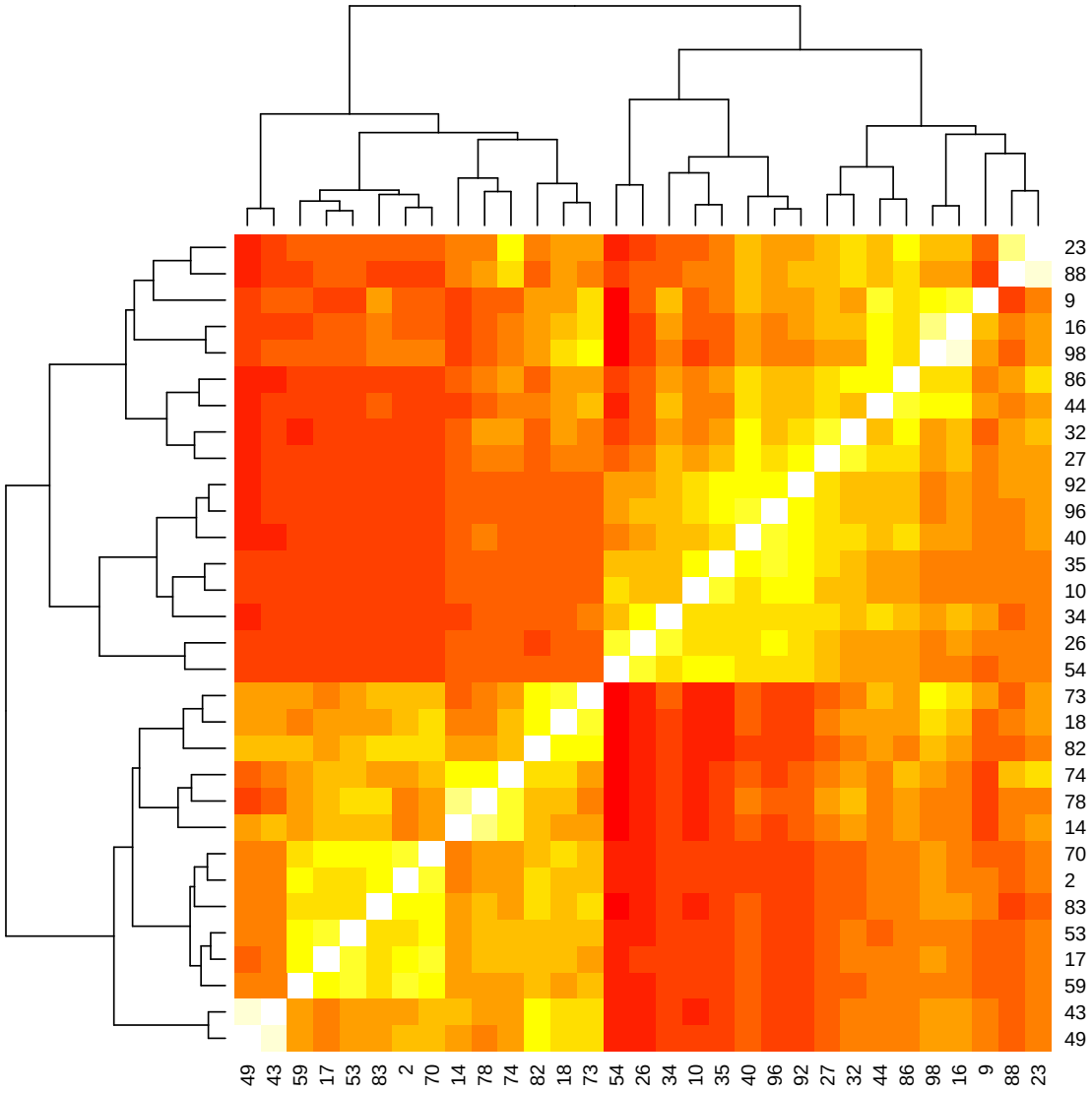


Figure 8.4: PRIM diversification clustering for assault