PREDICTING FIRE BRIGADE INCIDENTS

JEFFREY DE DEIJN (2516200)

Research Paper BA

MSc. Business Analytics
Department of Mathematics
Faculty of Sciences
VU Amsterdam

January 25, 2017

Jeffrey de Deijn: *Predicting fire brigade incidents*
Research Paper BA, © January 25, 2017

SUPERVISORS:
Prof. R.D. van der Mei (university)
G. Legemaate MSc (research company)

LOCATION:
VU Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081 HV Amsterdam

RESEARCH COMPANY:
Brandweer Amsterdam-Amstelland
Karspeldreef 16
1090 AD Amsterdam

PREFACE

The Research Paper BA is a compulsory part of the curriculum of the MSc. Business Analytics at the VU Amsterdam. The goal of this 6 ECTS (formally taking one month full-time) research is to demonstrate the student's ability to successfully go through the process of doing an individual research, to report on this in a clear and concise way and to clearly present it to all supervisors and expert managers.

Thanks to my supervisor from the VU, Prof. Rob van der Mei, I got into contact with Guido Legemaate of fire brigade Amsterdam-Amstelland. He is a pioneer on data intelligence and analytics in the fire brigade society of the Netherlands and he gave me the following research question:

Can we make a good forecast on the number of incidents that each fire station in Amsterdam-Amstelland has to handle? And in particular, can we make good use of weather forecasts in doing this?

I found this a very interesting topic to investigate and I am grateful to Guido that he entrusted me this task. Also, I want to thank him for supplying me such nice and clean data and for inviting me to the fire brigade headquarters to work on my research. Finally, I also want to give a big thanks to Rob for getting me in contact with Guido and for supporting me during the whole research with interesting conversations and good advices.

Jeffrey de Deijn, January 25, 2017

ABSTRACT

goal & approach The main goal is to make a forecast on the total number of fire trucks needed per day for each fire station. Most incidents need only one or just a few trucks, so we treat big incidents separately. If we define an incident as 'big' when it needs six or more trucks, then we can model the incident-arrival process as an (inhomogeneous) Poisson process: big incidents then occur with a slightly decreasing frequency (currently less than four times per week on average) and with independent and exponentially distributed interarrival times.

modelling Analysis on the small incidents shows that New Year's days should be treated separately as well. Together with some big storms, they form the main outliers in our dataset. For the remaining small incidents we need to correct for both the weekday and the time of the year (based on a Loess-smoothed function of the correction factors per weeknumber). After these corrections, we implement a linear regression model (LM), an LM with cross-term effects (GLM) and a random forest algorithm (RF), using four weather conditions (wind, temperature, rain and visibility) as explanatory variables. We create a different model for each *type-cluster* (we need clusters because some incident types do not occur frequently enough) and then divide the prediction among all fire stations based on average ratio's.

results & conclusions In general, GLM performs best in terms of (weighted) mean absolute percentage error ((w)MAPE), but RF is better in predicting 'busy' days. After some experimenting, it turns out that ensemble averaging ($\text{EA} = 0.2 \cdot \text{RF} + 0.8 \cdot \text{GLM}$) yields the best results ($\text{wMAPE} = 0.1860$ for daily totals). Rain and wind typically have a strong linear influence, while temperature mainly has non-linear influences. Besides some exceptions, most fire stations typically need only one fire truck. All stations have sufficient capacity.

recommendations The main improvements in future research can be made by investigating the approach of first aiming at a forecast per region. In addition, improvements may be made by reconsidering the clustering of incident types and by adding a trend correction before implementing RF.

CONTENTS

1 INTRODUCTION

1.1 FIRE BRIGADE AMSTERDAM-AMSTELLAND

Fire brigade Amsterdam-Amstelland helps improving and securing the safety in region Amsterdam-Amstelland. This region consists of six municipalities with a total of about one million inhabitants. The fire department is driven by 1100 men and women who intervene mainly in case of fire or other 'regular' incidents, but also in case of big disasters and crisis situation [1]. Amsterdam-Amstelland has nineteen fire stations with varying sizes; Each station has at least one fire engine available and some have even two of these and possibly also one rescue-, emergency- and/or water vehicle in addition. But what they all have in common is that all fire stations have people standing by 24 hours per day, ready to take action when they are needed.

1.2 PROBLEM AND GOAL

For the fire brigade, the question here is how many people have to be stand-by at each station at each time of the day. Having too many people stand-by costs a lot of money, but the fire brigade can't afford having too few people stand-by as well. How can this be optimized? Note that we cannot say anything about this without knowing how many incidents we may expect to happen. Therefore, the main question of this research is

Can we make a good forecast on the number of incidents that each fire station in Amsterdam-Amstelland has to handle?

Depending on the size and the type of an incident, the number of fire trucks (next: trucks) needed to handle this incident may vary from one or just a few to several dozens. Also the *type* of trucks needed varies among different incidents. However, in this study we will just treat any truck the same rather than making a distinction between different types.

1.2.1 *Previous research*

This question has been investigated in previous research, where the solutions are mainly based on multiplicative models containing correction factors for the weekday and the time of the year [2]. The main contribution of this research is to investigate the influence and predictive power of different weather conditions in the categories wind, temperature, rain and visibility. It would be great if the predictions can benefit from weather forecasts in the near future.

Previous attempts to make a separate prediction per district (27 in total) encountered the problem that there is too little data left after splitting it like this. This is due to the fact that each fire station has to deal with less than two incidents per day on average. Also in this research, dealing with the low number of incidents will be a challenging factor.

1.3 STRUCTURE OF THE PAPER

In the following, we will start by exploring the incident data supplied by the fire brigade. In Chapter 2 we will make a distinction between big and small incidents and discuss how we can handle the big ones. Chapter 3 then includes an analysis of the small incidents. Here, we first pay special attention to outliers in the data, after which we deepen into analysing trends, seasonal patterns and the influence of several weather conditions. Chapter 4 starts with a discussion on the different type of incidents and how we can use these in our model. Four different models for forecasting the daily number of trucks needed per fire station will be implemented and their results will be given along the way. Finally, Chapter 5 introduces a different approach, opening doors for future research and Chapter 6 gives some conclusions and recommendations for fire brigade Amsterdam-Amstelland. All analysis, modelling and forecasting in this research have been done in *R*.

2

BIG INCIDENTS

Incident data The available data contains one row for each incident that happened in region Amsterdam-Amstelland from January 2008 up until April 2016 (exactly 100 months, more than eight years). The most interesting information includes the incident's start- and end time, location and type as well as the concerned fire station and, last but not least, the number of fire trucks used.

The first choice we have to make is whether we want to make a forecast based on the number of incidents or on the total number of trucks used for these incidents. Since the size of incidents also matters for the amount of people you need, the most obvious choice here is to forecast the number of trucks needed. Hence, this is the first attribute to focus on. Figure 2.1 shows a histogram of the number of trucks per incident.

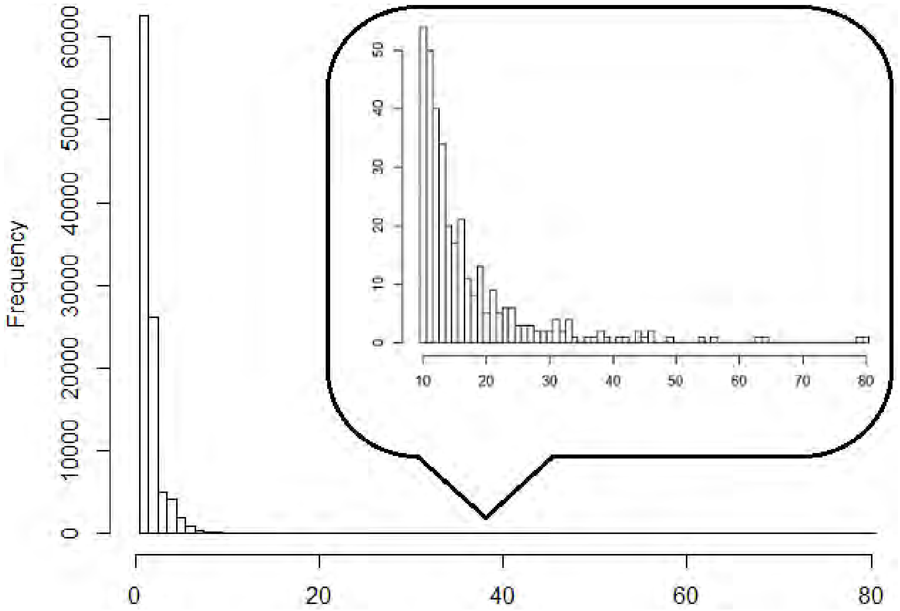


Figure 2.1: Histogram of the number of fire trucks per incident, including a zoom-in on the incidents with more than ten trucks.

2.1 DEFINING 'BIG'

As we can see, the vast majority of the incidents require only one or otherwise just a few trucks. Therefore, the question arises if it is convenient to make a difference between 'big' and 'small' incidents. And if so, how can we determine the most suitable boundary? In order to explore the characteristics of 'big' incidents, suppose for now that an incident is considered big if it requires at least 11 trucks (so that we get the zoomed-in histogram of Figure 2.1). Looking at the frequencies, such incidents seldom occur. More specifically, only 0.3% of the incidents can then be called big, which comes down to about two big incidents every three weeks on average. Incidents with more than 50 trucks are rare. The biggest incident in the dataset (in terms of how many trucks are used) corresponds to a big fire at an industrial terrain in Amsterdam-Noord. The fire started at Leenbakker, hit the surrounding buildings and 80 trucks were needed to control it. Actually, over 80% of the incidents with more than ten trucks were big fires. Incidents like this are due to coincidents that are impossible to predict. Specifically, they do not rely on bad weather conditions or a particular time of the year, as for example with forest fires in countries with a tropical climate. This arouses the expectation that big incidents can be modelled as a Poisson process. In order to check if this is indeed valid, we have to check if the interarrival times (the time duration between consecutive big incidents)

1. have an exponential distribution;
2. and are independent in time.

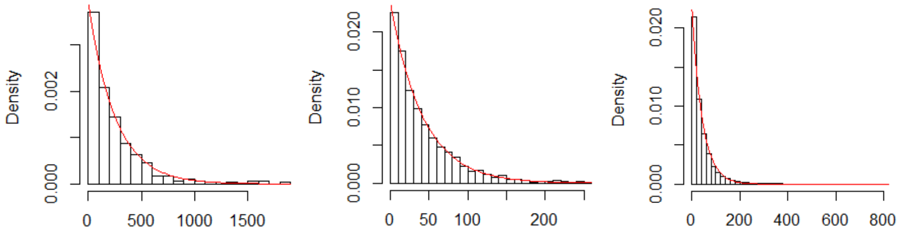
Check 1: exponentiality

We see in Figure 2.2a that the interarrival times seem to resemble the fitted exponential density very well. In fact, the Kolmogorov-Smirnov (KS) test does not reject that it is exponential (approximate p-value $= 0.529$[1]). This indicates that we can indeed assume the interarrival times to be exponential.

Check 2: independency

The Pearson's product-moment correlation test does not reject that the true correlation between the interarrival times and their order statistics is zero (p-value $= 0.9709$, sample correlation $= -0.002$). To

[1] This p-value is not computed in an exact way by using the regular one-sample KS-test. Since the exponential distribution to which it is compared has been fitted to the data, the resulting p-value would then be too high. Instead we have to use bootstrapping methods to approximate the correct p-value by simulation.



(a) Big when $\geqslant$11 trucks (b) Big when $\geqslant$6 trucks (c) Big when $\geqslant$1 trucks

Figure 2.2: Histogram of the interarrival times between big incidents where 'big' is defined differently in each plot. The red lines represent maximum-likelihood fitted exponential densities.

illustrate this, Figure 2.3 shows that there does not seem to be any change in trend. However, there is a significant 'clustering' of interarrival times (p-value $< 10^{-15}$), meaning that two consecutive interarrival times typically deviate less from each other than when they are farther apart in time. But, since the sample correlation is only 0.031, we can safely assume that also the independency holds.

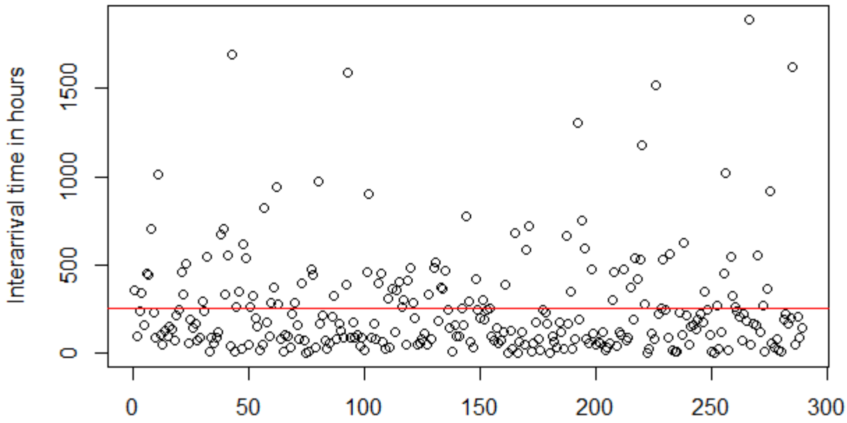


Figure 2.3: Plot of the interarrival times between 'big' incidents (with at least 11 trucks needed). The linear trend line is shown in red.

Conclusion 2.1 *If we define incidents as 'big' when the assistance of at least 11 trucks is required, then we can model the occurrence of big incidents as a Poisson process.*

2.2 OPTIMIZING THE DEFINITION

What if we now redefine when we call an incident 'big'? If we shift the boundary, the histogram of the interarrival times keeps looking exponential at first sight, even if we define all incidents as being 'big' (see Figure 2.2c). However, the KS-test allows us to make a clear distinction: if we define an incident as 'big' when at least six trucks are used, then the KS-test does not reject exponentiality of the interarrival times (approximate p-value = 0.429). This also holds for any boundary above six. However, for lower boundaries the KS-test doubts (or rejects) this exponentiality (approximate p-value = 0.073 and 0.002 for 'big' when at least five respectively four trucks used). Hence, according to this we wish to set the boundary at six.

Check independency

The remaining question now is whether the independency still holds in this case? Concerning the clustering issue, we can use similar arguments as before (sample correlation = 0.051). However, Figure 2.4 shows that the interarrival times seem to increase in time now we added incidents with six to ten trucks. More specifically, the trend line indicates that the average interarrival time has increased by 17.35 hours over the full time span of the data, which is an increase of over 50%! Apparently, the fire brigade succeeded in improving the fire safety such that those intermediate incidents occur much less, while the real big incidents are still hard to prevent. If we now correct for this trend, the dependency becomes insignificant again (p-value = 0.5404), while the exponentiality is still not rejected as well (approximate p-value = 0.462).

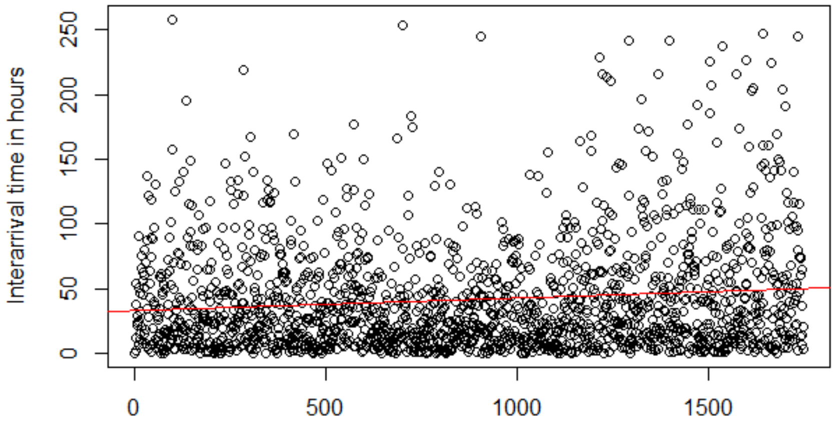


Figure 2.4: Plot of the interarrival times between 'big' incidents (with at least six trucks needed). The linear trend line is shown in red.

Conclusion 2.2 *We will define incidents as 'big' when the assistance of at least six trucks is required. This is the boundary case for which we can still model the occurrence of big incidents as an (inhomogeneous) Poisson process. Consequently, all incidents that require at most five trucks will be referred to as 'small'.*

2.3 DEALING WITH BIG INCIDENTS

Using the definition from Conclusion 2.2, 1.72% of the incidents are big (for histogram of interarrival times, see Figure 2.2b), which comes down to around four big incidents every week (still most of these are big fires). One may say that this is quite frequent, but their occurrence is very hard to predict. Therefore we probably cannot do much better than modelling them as a Poisson process. If we want to account for the trend found in Figure 2.4, we have to use an *inhomogeneous* Poisson process with a decreasing rate. However, the change is such slow that using a regular homogeneous Poisson process is fine as well. We can use this if we want to do simulations. Here we can also use the distribution of big incidents amongst all fire stations as illustrated in Figure 2.5.

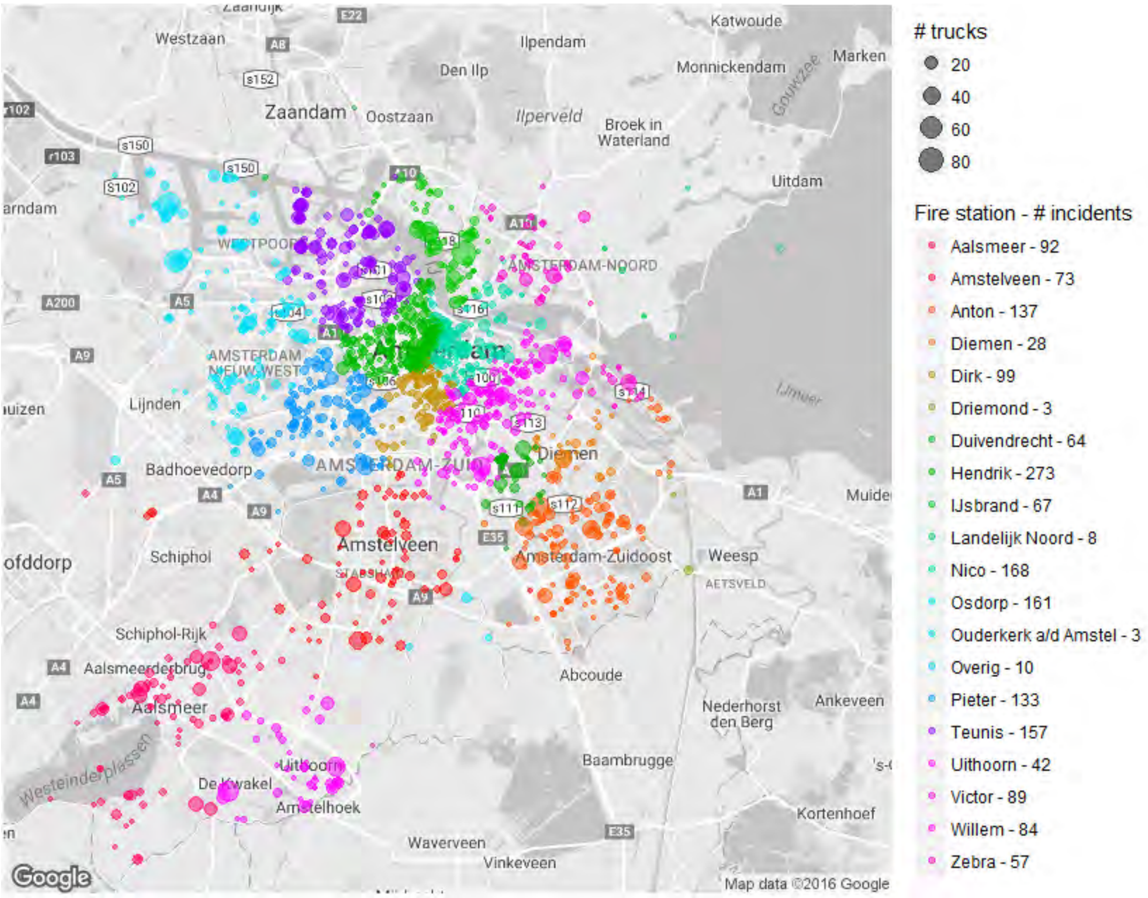


Figure 2.5: Plot of the locations of all big incidents. The size of a point corresponds with the size of the incident and its colour with the fire station responsible for tackling it. The ten 'Overig' (English: 'other') big incidents fall outside the plotted area (two of them in France, one in Purmerend, one in Velsen and six in Soest).

2.3.1 *Results for the final model*

For computing confidence intervals for the number of trucks needed per station, we can only use the information given in Table 2.1. It will probably be a good idea for each station to account for big incidents by adding up this 95%-quantile (or similar quantiles) to the number of fire trucks needed for small incidents. Note that these quantiles are just an indication of what to add up for each station. Finally, one may choose to use quantiles that decrease a little every day along with the decreasing rate of the Poisson process. The difference in results will be very small though.

Table 2.1: Statistics of the number of trucks used for big incidents per day per fire station.

Station	**% Trucks**	**Avg**	**SD**	**95%-quantile**
Hendrik	14.8%	0.74	2.9	2.82
Teunis	10.0%	0.50	2.5	1.91
Osdorp	9.6%	0.48	2.8	1.82
Nico	8.9%	0.45	2.1	1.69
Anton	7.8%	0.39	2.3	1.48
Pieter	7.6%	0.38	2.2	1.45
Aalsmeer	5.6%	0.28	1.9	1.06
Dirk	5.0%	0.25	1.5	0.96
IJsbrand	4.9%	0.25	2.4	0.93
Victor	4.9%	0.24	1.7	0.93
Amstelveen	4.4%	0.22	1.7	0.84
Willem	4.4%	0.22	1.6	0.84
Duivendrecht	3.4%	0.17	1.3	0.64
Uithoorn	3.0%	0.15	1.6	0.57
Zebra	2.9%	0.15	1.2	0.56
Diemen	1.5%	0.07	0.9	0.28
Overig	0.5%	0.03	0.5	0.10
Landelijk Noord	0.3%	0.02	0.4	0.07
Ouderkerk aan de Amstel	0.2%	0.01	0.3	0.04
Driemond	0.1%	0.01	0.2	0.03
Total	100%	5.0	8.0	19

Conclusion 2.3 *In our final model, we will use quantiles such as the ones in Table 2.1 in order to determine the capacity needed for big incidents.*

3 SMALL INCIDENTS

3.1 NEW YEAR'S EVE

The small incidents are probably easier to predict, since for example bad weather conditions often cause many *small* incidents to happen (like fallen trees, water damage or police/ambulance assistance at traffic accidents). To illustrate this, we take a closer look at the outliers observed in Figure 3.1. It is notable that there are always (17 days in total, either December 31 or January 1) many incidents around New Year's Eve. Figure 3.2 indicates that this is mainly caused by accidents involving firework. These conditions do not occur in the rest of the year, so it seems logical to analyse these days separately. For this purpose, we can take the regular forecast for these days (based on weekday/weather/etc.) and subtract this from the number of trucks per hour on the 17 'New Year's days'. If we then assume that the resulting data concerns firework incidents, we can analyse the characteristics of such incidents (which may be very different from that of the 'average' incident). For instance, firework incidents may be more/less prone to bad weather conditions than other incidents. We will check this later, when the regular forecasts are available. In the following, all big incidents as well as all incidents on December 31 and January 1 are omitted.

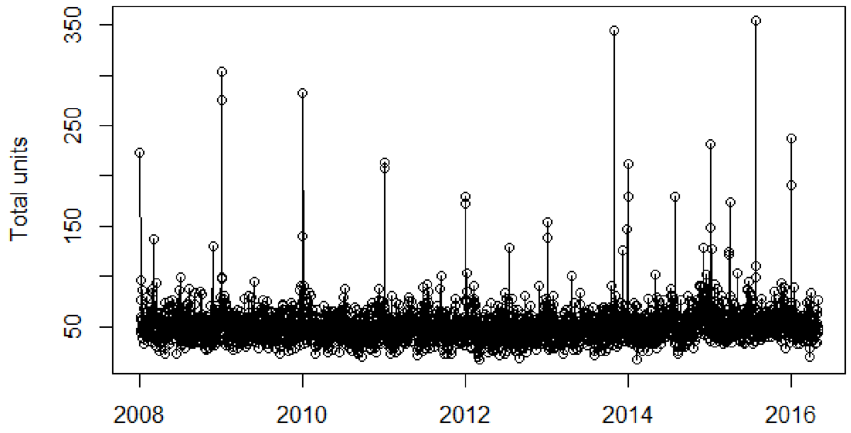
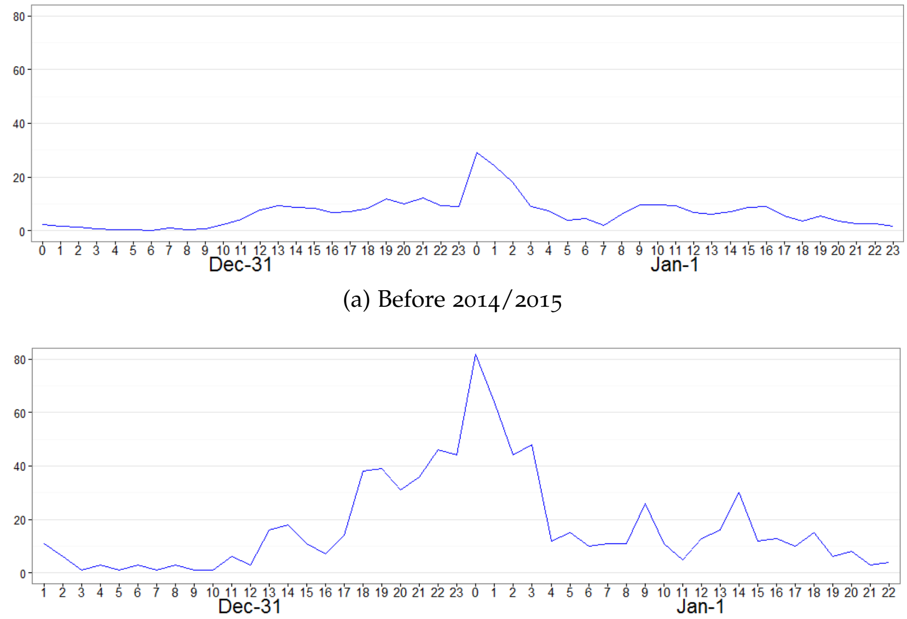


Figure 3.1: Plot of the total number of trucks used for small incidents per day. Almost 99% of the days less than 100 trucks were needed, but there were some big outliers. On the other hand, there were no days on with extraordinary few trucks were used (the minimum is 17). Most of the days around 50 trucks were needed.



(a) Before 2014/2015

(b) From 2014/2015

Figure 3.2: Plot of the average number of trucks used for small incidents per hour around New Year's Eve. The relation with firework is obvious; Until New Year's Eve 2014/2015, it was allowed to use firework from 10am on December 31 until 2am on January 1. Since then the starting time changed to 6pm. Also, the difference between before and after the change of regulations is striking, but it is unclear whether this is just a matter of chance or not.

Conclusion 3.1 *Due to firework incidents, there are extremely many incidents on December 31 and January 1 and also the daily pattern of incidents differs on these days. It is therefore wise to treat these days separately from the rest.*

3.2 OTHER OUTLIERS

To begin further analysis, it is important to note that not all outliers in Figure 3.1 are New Year's days. However, the only five days that could compete with these outliers (i.e. with more than 138 trucks used for small incidents) are days with severe weather conditions. On the five peak days in Table 3.1, only 0.46% (instead of the average 1.72%) of the incidents is big. There is no reason to assume that the occurrence rate of big incidents is lower on these days. So apparently - if the difference is not just a matter of chance - there are certain circumstances that cause many, especially small incidents to happen.
Therefore the question arises whether we can predict incidents of a certain size? For instance, do incidents with one truck have different

Table 3.1: The weather conditions on the five days that could compete with the New Year's days in terms of number of trucks used. Apparently, wind and rainfall can have a big impact.

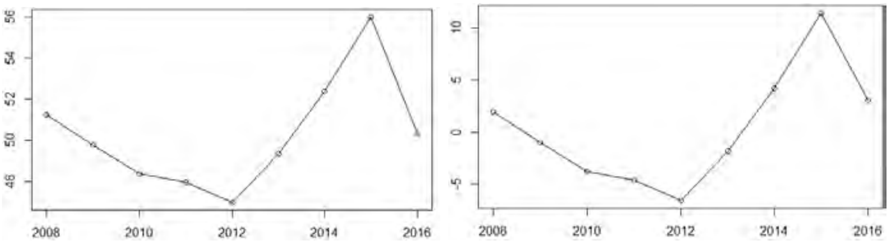
		Highest windspeed (km/h)		Total rainfall (mm)	
Date (day)	Trucks	Worst hour	Overall	Worst hour	Overall
28/10/2013 (Mon.)	345	79.2	111.6	3	10.7
24/12/2013 (Tue.)	147	64.8	111.6	2.3	6.8
28/07/2014 (Mon.)	179	28.8	43.2	12.6	60.5
31/03/2015 (Tue.)	174	64.8	100.8	3.8	7.9
25/07/2015 (Sat.)	355	72	100.8	5.8	19.7

characteristics than those with two trucks? This does not sound very promising, but it becomes better if we look at it from another perspective. It may be the case that a certain *type* of incident typically needs only one or two trucks while another often needs a few. For instance, in 96.6% of the cases only one truck is needed to free a person that is stuck in an elevator. Also, in 96.1% of the cases exactly two trucks are needed for an 'Afhijsen spoed' incident and in 85.4% of the cases at least three trucks are needed for an inside fire ('Binnenbrand'). Since different types of incidents may very well have different characteristics, we can probably distinguish them to improve our forecast.

Conclusion 3.2 *The outliers in Table 3.1 indicate that the weather probably is important in predicting especially certain types of incidents.*

3.3 STRUCTURAL TRENDS

Before we dig into incident types and the weather, let's start with an overall analysis of the small incidents. First, Figure 3.3b does not show a clear structural trend. Although the pattern is remarkable, no explanation for this can be found in the yearly development per month in Figure A.2.



(a) In average trucks per day

(b) In % trucks above average

Figure 3.3: The yearly development of the average trucks per day. In calculating the percentage for 2016, we correct for only having data up to and including April.

July may be the only exception, having a steep growth in the last two years. This growth is unlikely to continue, but it may partly be explained by the current climate change. Lenderink et al. [4] conclude that the global warming can explain the increase of extreme rainfall, especially during the summer. In this context, recall the heavy rainfall in especially the two days in July in Table 3.1 and also note that all weather outliers occurred in the last years of our dataset. In addition, Figure 3.4 illustrates the positive trend in the amount of rainfall in July. Also Pearson's product-moment correlation test confirms that the true correlation between rainfall in July and time (in years) is positive (p-value = 0.007; sample correlation = 0.364). These are all signs that the fire brigade may consider accounting for the increasing chance of extreme weather conditions by acquiring enough fire trucks across the board to be able to encounter such outliers as in Table 3.1.

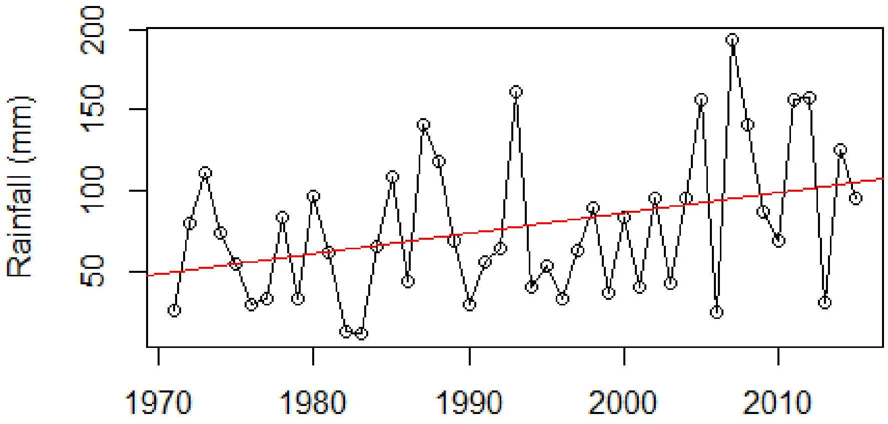


Figure 3.4: Plot of the yearly rainfall in July. The linear trend line (in red) indicates a growth of 1.259mm per year.

The argument of climate change forces us to double-check if we really don't need to account for this in our model. Figure 3.5 shows that the empirical linear trend line indeed has a positive slope. And again, the Pearson correlation test confirms that also the true slope of this line is positive (p-value $< 10^{-5}$ and sample correlation between trucks per day and time in days is 0.081). Unless the growth is very weak, we may therefore include a trend component in our models. Note that only 16.4% of the small incidents is due to fire, so we cannot conclude that the actions to improve fire safety failed for small fires. In fact, as for the big fires (see Section 2.1), the number of trucks used for small fires decreased, namely by 16.7% over the regarded time span.

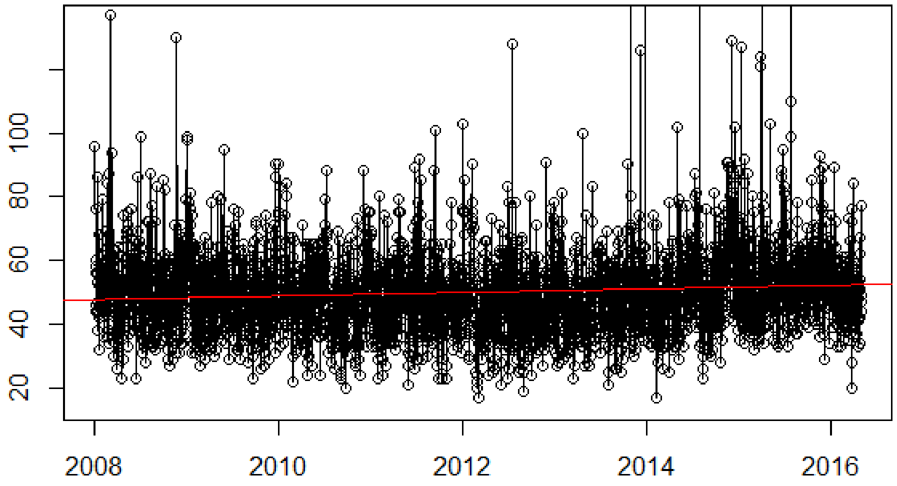
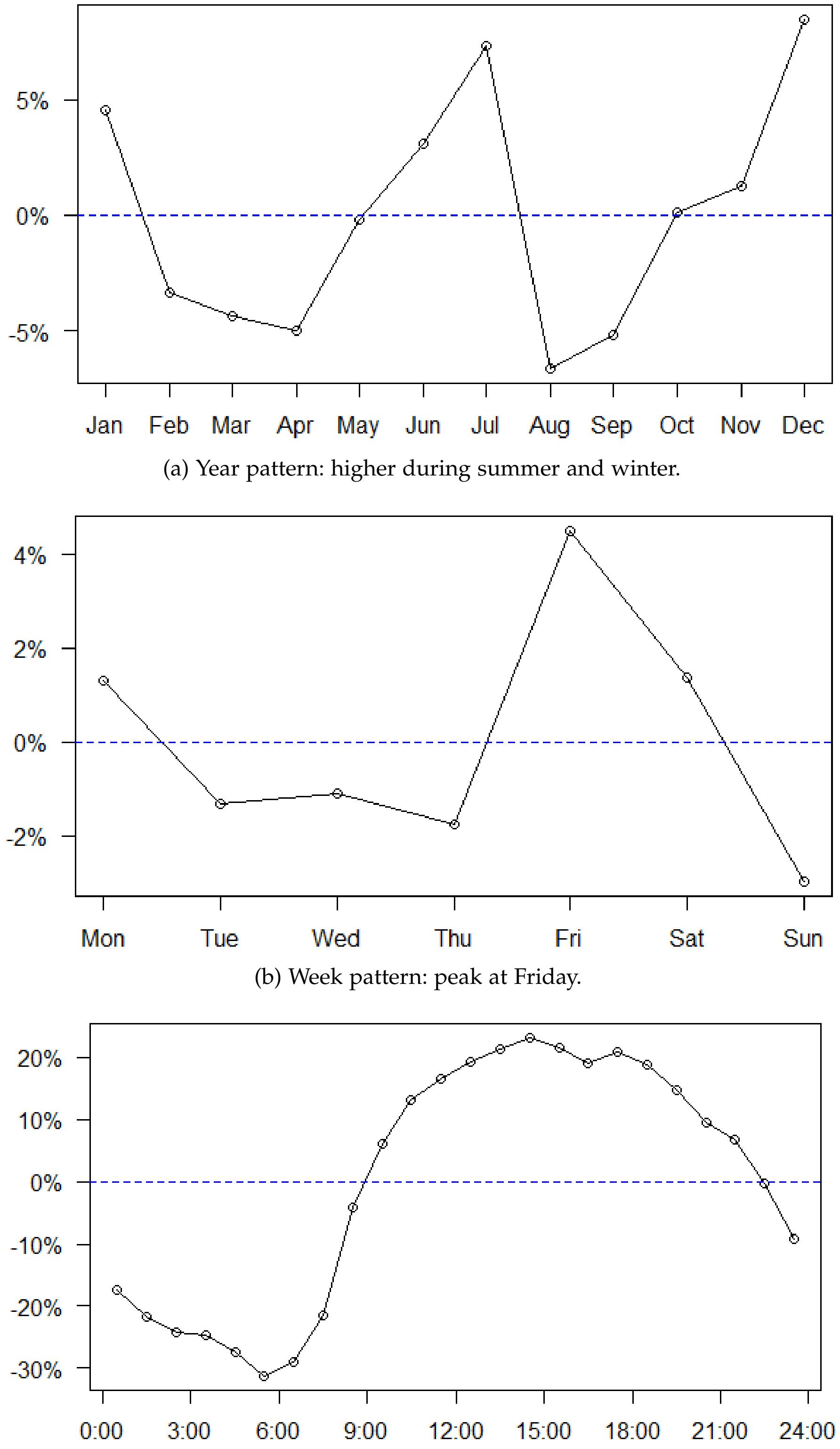


Figure 3.5: Plot of the total trucks per day (excluding the New Year's days). In order to see more detail, the peaks of the five outliers from Table 3.1 are not shown. The linear trend line (in red) indicates a growth of 0.001465 trucks per day (~0.535 trucks per year).

Conclusion 3.3 *There is a very weakly increasing trend in the number of incidents. We may consider accounting for this in our models, but it will probably not have a major effect.*

3.4 SEASONAL PATTERNS

Besides a trend, the number of trucks also shows specific patterns throughout each year, week and day. The plots in Figure 3.6 illustrate this. We will have to include all these patterns in our model later on. For now, we must note that the day pattern in Figure 3.6c is most striking. It is important to know if this pattern is always and everywhere the same. Therefore, we plotted the day pattern per weekday and region (cluster in this case) in Figure A.3. From this we can conclude that no problems arise here. We can just use the average day pattern in our model for all days and regions.



(a) Year pattern: higher during summer and winter.

(b) Week pattern: peak at Friday.

(c) Day pattern: low at night, high at midday.

Figure 3.6: Seasonal patterns of trucks used for small incidents (New Year's days excluded). The given percentages represent relative differences with respect to the average (in blue).

The pattern in Figure 3.6a can be included in the model in a more subtle way than taking factors per month. The problem here is that, for instance, the differences between the beginning and end of January are considerable. Instead of factors per month, we can therefore better use factors per week as shown in black in Figure 3.7. However, the degree of fluctuation of this graph does not agree with our expectations. We would expect that the real pattern is much smoother, so therefore the red graph represents a smoothed variant. This already looks much more realistic, so this is the one we will use in our model.

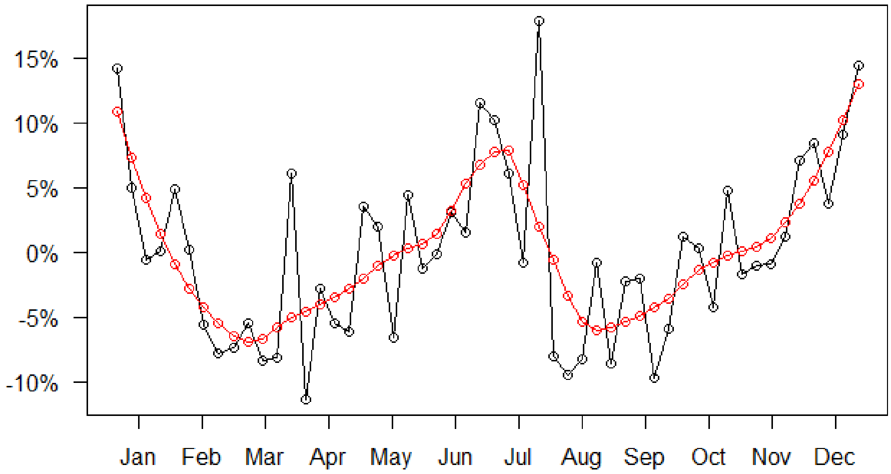


Figure 3.7: The year pattern per week (in black) together with its Loess-smoothed variant ($\alpha = 0.3$). The week numbers used are slightly different than the regular week numbers to avoid problems with week 53.

Interpreting the patterns

Our common sense tells us that everything happens for a reason. Small fluctuations can occur randomly, but not the general patterns. The challenge now is therefore to explain why certain patterns look the way they do. The pattern in Figure 3.6c is still easy. It illustrates the activity cycle that an average person goes through every day of the week (recall Figure A.3a). This pattern is also nearly the same at any time of the year, even during the summer holidays (see Figure A.4a). And where there is more activity, there is more risk for incidents to happen. No further causes have to be searched for here. However, the bigger the time span becomes, the more likely it is that circumstances change. This makes it harder to know which factors influence these patterns, and to what extent. For example, Figure A.4b shows that the week pattern from Figure 3.6b looks a little different throughout the year. The differences are not frightful, but how can we explain them? The weather cannot be blamed. Why would it be relatively bad on Saturdays in July? Although in July the fraction of incidents that is due to storm and water damage is somewhat higher (11.7% vs. 6.1%), the differences regarding the types of incidents are small. This is not surprising, since we have seen in Table 3.1) that the biggest outlier is a Saturday in July. This raises the question whether or not the differences we observe between the weekdays are just a matter of chance. Therefore, we use two-sample *t*-tests to test if the average number of trucks differs significantly on different weekdays. Using a significance level of 5%, these tests yield a significant difference only between Friday and the four days that are below average. However, the differences are too big to ignore, so we will include them in our model. Also because the differences are even bigger when we consider incidents of a particular type. For example, an inside fire ('Binnenbrand') will occur 13.3% more often during the weekend, because more people are at their homes then. In this case, the weekday factors will probably be very important.

Conclusion 3.4 *The day pattern is quite standard, so we will just make forecasts per day. The week pattern differs per type of incident and has to be included in our model. The year pattern can best be corrected for by using a Loess-smoothed function over the factors per week.*

3.5 WEATHER INFLUENCES

Besides the time dependent components, we want to know which weather variables we must include in our model. Therefore, we use again the Pearson correlation test to determine which weather conditions have a significant influence on the number of trucks we need for small incidents. The results of these tests are summarized in Table 3.2. We can see from this that the minimum visibility and the average temperature both have no significant (direct) influence. However, if we consider a variable indicating whether it was on average freezing on that day, then this does have predictive value. Obviously, we also have to include some variables indicating the amount of rainfall and wind. However, the variables within these categories are highly correlated (sample correlation around 0.9) and therefore we may exclude some of them to simplify our model. Probably, including the maximum wind gust and total rainfall will be satisfactory, but we will do some tests for this in the next chapter.

Table 3.2: Results of Pearson's product-moment correlation tests between some weather variables and the number of trucks used for small incidents per day. All tests are one-sided, except from the daily mean temperature, for which we had no clear hypothesis about the sign of the correlation.

Category	Variable	P-value	Sample correlation
Wind	Average windspeed (FG)	$< 10^{12}$	0.132
	Maximum hourly mean windspeed (FHX)	$< 10^{15}$	0.177
	Maximum wind gust (FXX)	$< 10^{15}$	0.189
Temperature	Average temperature (TG)	0.6897	0.007
	Boolean: 1 if average > 0 (TG>0)	$< 10^{8}$	0.105
Rainfall *	Rainfall duration (DR)	0.0004	0.061
	Total rainfall (RH)	$< 10^{15}$	0.151
	Maximum hourly rainfall (RHX)	$< 10^{12}$	0.132
Visibility **	Minimum visibility (VVN)	0.2217	-0.014
	Boolean: 1 if minimum $<$ 200m (VVN<2)	0.2893	0.010

** In 0.1 mm and -1 for <0.05 mm; ** On 0-89 scale, where 0: <100 m, 89: >70 km.*

Influences per incident type

The choice of which variables to include depends also on the characteristics of different types of incidents. For example, on rainy/windy days there will typically be less outside fires, but more incidents due to storm and water damage. Moreover, it may be the case that some incidents occur more often when it's cold and others when it's warm, balancing the overall effect of the average temperature. These type characterizations will turn out to be useful in modelling. We will come back to this in Section 4.1. A final note here is that the data from the New Year's days are excluded in the computations of Table 3.2. This causes only a tiny loss of data and enables us (later on) to analyse the weather influences on the firework incidents separately.

Conclusion 3.5 *Wind and rainfall have a clear positive correlation with the number of incidents. However, each type of incident responds differently to certain weather conditions. It is important to capture this in our model.*

4 FORECAST PER FIRE STATION

Before we start off modelling, let's have a short recap of what our objective is and what we have learned from the previous chapters. As our first approach, we will create a model that predicts directly the number of trucks that each fire station needs. In Chapter 2 we have seen that big incidents (with at least six trucks needed) are very hard to predict and that we can best model them by an (inhomogeneous) Poisson process. Then, Chapter 3 showed that the daily pattern of the number of trucks used for small incidents is quite standard. So if we know for some day how many trucks are needed in total, we can quite accurately extract from this how many trucks are needed per hour. Therefore, we will try to forecast the number of trucks needed per day per fire station. We will do this based on

1. the number of trucks that each fire station needed in the past;
2. the (expected) weather conditions;
3. and the characteristics of each type of incident.

4.1 TYPE CHARACTERISTICS

In total we have 29 different incident types in our dataset, some of which occur much more/less often than others. The question is whether they are all relevant enough to make a forecast for. In Table B.1 in the appendix, we see that some incident types do not occur frequently enough to be able to make a good forecast. This is not a big problem, since we can just ignore these incident types. They would hardly increase the predictions anyway. This holds at least for most types in the lower block of Table B.1. For types 23-26 in this table, we can quite confidently state that these types are mapped to other types from some date onwards by looking at the patterns in Figure A.1. Table 4.1 summarizes which transitions have been made. From this we can conclude that we can safely ignore all incident types in the lower block of Table B.1. Moreover, we have to take into account that there are some breaks in the data of the types in the lower block of Table 4.1. In order to prevent incorrect inference, we therefore can only use the data of incident type 'Hulpverlening algemeen', 'Assistentie Ambulance' and 'Assistentie Politie' from 2014/06 onwards, the data of incident type 'Reanimeren' from 2012/07 onwards, and the data of incident type 'Hulpverlening Dieren' from 2012/03 onwards.

Table 4.1: Overview of the type transitions that (probably) have been done.

From	To	When	Reversed
Interregionale bijstand	Assistentie ambulance	2012/03	-
Hulpverlening algemeen dieren	Hulpverlening dieren	2012/03	-
Afhijsen spoed	Hulpverlening algemeen	2012/03	-
Beknelling / bevrijding	Hulpverlening algemeen	2013/06	-
Reanimeren	Assistentie ambulance	2012/03	2012/07
Assistentie ambulance (partly)	Hulpverlening algemeen	2013/06	2014/06
Assistentie politie (partly)	Hulpverlening algemeen	2013/06	2014/06

Type-clustering

We now still have some incident types that we don't want to ignore, but which do not occur on a daily basis. Therefore, we may think of a way to cluster the incident types such that each 'type-cluster' occurs frequently enough to make a reasonable prediction for it. Since we want to predict on weather data, it may be a good idea to base the clustering on this. Two types are put in the same cluster (manually) when they have similar correlation with respect to the weather variables. After some experimentation, the most suitable clustering seems to be such as in Table 4.2. For each cluster, we also show the correlation with respect to one variable of each four weather categories (recall Table 3.2). We choose in each case the variable to which the cluster has on average the highest absolute correlation (no variable is chosen when this highest correlation is less than 2.5%). Looking to these correlations in detail, we can conclude that these are often in line with our expectations. For instance, high windspeed and rainfall obviously increase the number of incidents due to 'storm and water damage' (type-cluster 9) and decrease the likelihood of 'outside fires' occurring (type-cluster 1).

Conclusion 4.1 *Table 4.2 gives a type-clustering such that each type-cluster occurs on a daily basis and contains types with similar characteristics with respect to the weather. Due to the transitions in Table 4.1, we can use the data of type-cluster 2 only from 2012/03 onwards and the data of type-clusters 6, 7 and 8 from 2014/06 onwards.*

Table 4.2: Type clustering and their correlation with respect to windspeed, temperature, rainfall and visibility. In addition, the average number of small incidents per day is given for each cluster (note: always > 1). Here, we have taken into account the date from which we can use the data of each type-cluster. New Year's days are excluded from this analysis.

Cluster	Type	Wind	Temp.	Rain	Visib.	# p/day
1	Buitenbrand	-0.135	0.09	-0.193	0.075	3.46
2	Dier te water	-0.088	0.134	-0.058	0.013	1.65
	Hulpverlening dieren	-0.072	0.129	-0.088	0.069	
	Persoon te water	-0.041	0.056	-0.023	0.009	
	Buitensluiting	-0.006	0.159	-0.043	0.062	
3	Meten / overlast / verontreiniging	-	-0.228	0.038	-0.111	2.52
4	Liftopsluiting	-	-0.088	0.021	-0.015	8.16
	OMS / automatische melding	-	-0.069	0.051	-0.037	
5	Brandgerucht / nacontrole	-	-0.103	-	-	3.57
	Binnenbrand	-	-0.038	-	-	
	Hulpverlening water algemeen	-	-0.019	-	-	
6	Assistentie politie	0.048	-0.062	0.026	-	1.34
7	Assistentie ambulance	-	-0.065	-	-0.039	8.55
	Voertuig te water	-	-0.042	-	-0.025	
	Reanimeren	-	-0.086	-	-0.008	
8	Hulpverlening algemeen	0.063	0.079	0.057	0.052	2.28
9	Storm en waterschade	0.319	0.028	0.279	-	2.10

4.2 SMALL INCIDENTS: MODELLING AND RESULTS

The first intention was to predict the number of trucks needed for small incidents separately for each fire station/type-cluster combination. However, for most of these combinations it turned out that there are way too few incidents to make an accurate forecast. This is already implied by the last column of Table 4.2 and confirmed by Table B.2 in the appendix. We will therefore just make a separate forecast for each type-cluster, and then share this forecast among all fire stations according to the percentages in Table B.3. We don't lose too much information with this, since the characteristics of type-clusters do not differ much between different fire stations. If for fire station X some type-cluster occurs more on Mondays or when it rains, then this is very likely to hold also for other fire stations. So in short, we will estimate, for each type-cluster t, a model that predicts the number of trucks used for *small* incidents $y_{t,d}$ on date d, i.e.

$$y_{t,d} = f_{t,d} g_{t,d} x_{t,d}. \tag{1}$$

Here, $f_{t,d}$ is a correction factor for the week number based on a Loess-smoothed function as in Figure 3.7, and $g_{t,d}$ is a weekday factor as in Figure 3.6b. Both are computed separately for each type-cluster. Finally, the term $x_{t,d}$ contains all remaining information. This includes the average level, dependencies on the weather, a possible trend and dependencies on all other variables that we are currently not considering, but which do exist in reality.

We will consider three different ways of modelling $x_{t,d}$, after which we compare their performances. We will do this by splitting the available data into a training set (on which we estimate our models) and a test set (for which we make forecasts using the estimated models). Recall that for some type-clusters we can only use the data from 2014/06 onwards and that we have data up until 2016/04. Since it is desirable to have at least one year of data in our training set, we take here all data up until 2015/06 [1]. Note that for the type-clusters without any restriction, the training set contains the data from 2008/01 up until 2015/06. Therefore, the quality of the predictions for these type-clusters may be a lot better then for the type-clusters for which the training set only contains one year of data.

Performance measure

The test sets contain all data from 2015/07 onwards. This holds for all type-clusters, so all test sets contain exactly nine months of data and the quality of the forecasts can therefore be compared easily. We will measure the quality of a forecast using the Mean Absolute Percentage Error,

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \overset{(y_t \geqslant 0)}{=} \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t}, \tag{2}$$

as well as its weighted version, i.e.

$$\text{wMAPE} = \frac{\sum_{t=1}^{n}\frac{|y_t - \hat{y}_t|}{y_t}y_t}{\sum_{t=1}^{n} y_t} = \frac{\sum_{t=1}^{n}|y_t - \hat{y}_t|}{\sum_{t=1}^{n} y_t}. \tag{3}$$

Here, y_t is the true value in time period t and $\hat{y}_t$ is the prediction. Note that the last equality in Equation 3 does not hold when $y_t = 0$ since division by zero is not defined. However, it is common practice to compute it like this, so we will do this as well.

Conclusion 4.2 *After a correction for the weeknumber and weekday, we will use four different methods for modelling the daily number of trucks needed for small incidents. Then, we will add up the results for the big incidents (recall Table 2.1), define a rule to determine the needed capacity per day for each fire station, and also analyse the firework incidents.*

[1] After documenting all results I noticed that I could have taken 2015/05 here, but since this is not really a problem anyway, I left it like this.

4.2.1 *Linear regression model*

The first attempt to model $x_{t,d}$ from Equation 1 is by means of the linear regression model (LM)

$$\begin{aligned} x_{t,d} = \beta_0 &+ \beta_1 \cdot d \\ &+ \beta_2 \cdot \text{windspeed}_d \\ &+ \beta_3 \cdot \text{temperature}_d \\ &+ \beta_4 \cdot \text{rainfall}_d \\ &+ \beta_5 \cdot \text{visibility}_d \\ &+ \epsilon_{t,d}, \end{aligned} \tag{4}$$

where $\epsilon_{t,d}$ is assumed to have expectation zero and some finite variance. Its distribution does not resemble a normal distribution, because it has a fat right tail. It therefore looks more like a shifted log-normal distribution. However, we will not deepen into this, because we actually do not need a fitted distribution. Later on, we will see that we can compute prediction intervals for the test set using the empirical distribution of the residuals in the training set.

Note that this model includes an intercept (β_0), a linear trend ($\beta_1 \cdot d$) and (at most) four weather variables. For each type-cluster, we will use the same weather variables as in Table 4.2. This implies that we exclude those variables that have almost no effect on that type-cluster.

Model estimation

It will be interesting to see the estimated parameters of the model. The question is whether the weather variables that have high correlation to some type-cluster also have high predictive power. We can say that a variable has predictive power if its estimated parameter is significantly bigger/smaller than zero. Therefore, we will conduct a two-sided *t*-test to test the null hypothesis $H_0 : \beta_j = 0$ against the alternative $H_1 : \beta_j \neq 0$. The smaller the p-value of this test, the more certain we are that β_j is truly unequal to zero, or equivalently, that variable j has predictive power.

The estimated parameter values can be found in Table B.4 in the appendix. However, instead of the exact values, it is more interesting to see how significant a parameter is on a 1 to 5 scale, as in Table 4.3. Here, we assign 1 when the $\text{p-value} < 0.001$ (very significant) until 5 when the $\text{p-value} \geqslant 0.1$ (not significant). Comparing this to Table 4.2, we observe that when a weather variable has significant predictive power for some type-cluster, then their mutual correlation is relatively high as well. This is a nice result, but unfortunately the reverse is not true. For instance, type-cluster 3 is highly correlated with one of the temperature variables, but this variable does not have predictive power for this type-cluster, which is surprising.

Table 4.3: Significance on a 1-5 scale of the parameters of the estimated linear models. Remember that a different model is estimated for each type-cluster. The scale is based on the p-value of two-sided *t*-tests.

	Type-cluster									
Variable	1	2	3	4	5	6	7	8	9	Avg
Intercept	1	1	1	1	1	4	1	1	1	1.33
Trend	1	5	1	4	3	5	5	5	5	3.78
Windspeed	1	5	5	3	5	5	5	5	1	3.89
Temperature	3	4	5	1	2	5	5	5	5	3.89
Rainfall	1	3	5	5	5	5	5	4	1	3.78
Visibility	5	4	5	4	5	5	5	5	5	4.78
Scaling: 1: p<0.001, 2: p<0.01, 3: p<0.05, 4: p<0.1, 5: p<1										

If we look at Table 4.3 in more detail, it stands out that several type-clusters have no weather variables with significant predictive power. Opposed to type-cluster 3, this is not surprising for type-clusters 6 and 7, since their correlations to the weather variables are relatively low as well. On the other hand, type-clusters 1 and 9 are well predicted by the amount of wind and rainfall, which is intuitively explainable as well (recall the example at the end of Section 4.1).

Performance per type-cluster

Now we want to know of course how good this model is at predicting the required number of trucks. Since we created a different model for each type-cluster it is at first interesting to make a prediction per day for each type-cluster. For some days in the test set, there were zero small incidents in reality, so note that the MAPE does not work here. Table B.8 shows the wMAPE for each type-cluster. If we compare this to Table 4.3, we conclude that it is in general not the case that the more weather variables with predictive power a type-cluster has, the more predictable it is. For instance, type-cluster 7 it best predictable, while it has no weather variables with significant predictive power. On the other hand, we know that wind and rain have a huge effect on the number of incidents of type 'storm and water damage' (type-cluster 9). The explanation for this can be found by looking at the amount of variation in the real data. We therefore added to Table B.8 a column with the Coefficient of Variation (CoV; the standard deviation divided by the average). We observe that the number of trucks used for small incidents is most stable for type-cluster 7, which is why this type-cluster is easier to predict. The CoV of type-cluster 9 is maybe even more striking. We have already seen that the biggest outliers in our dataset are due to severe weather conditions (recall Table 3.1). Most incidents on these days are from type-cluster 9, so it is no surprise that its CoV is so high. If we correct the wMAPE by dividing it by the CoV, we actually see that we are doing a very good job in predicting type-cluster 9. A much bigger part of its variation can be explained than for the other type-clusters.

Performance per fire station

Instead of a prediction per type-cluster, it may be more interesting for the fire brigade to look at the prediction per fire station in Table B.9. At first sight, the wMAPE does not look very good for many fire stations. However, if we compare it to the average number of trucks used for small incidents per day for each fire station, we can conclude that it is very hard to accurately predict the number of trucks when the average is very low. For instance, when we predict 2 when it turns out to be 1 in reality, then we already make an error of 100%. However, the question is how bad this really is in this case. In any case, it is unlikely that an error of one truck will have a disastrous impact in real life. Moreover, when the average is such low as for some fire stations, then the majority of the days will have no incidents at all. Since we look at the *weighted* MAPE, these days have no weight. In other words, on these days we only get punished for making an error, but we are not rewarded at all for being close (recall Equation 3). In order to clarify the relationship with the average number of trucks used for small incidents per day, Figure 4.1 shows a scatter plot where each point represents a fire station. Obviously, the wMAPE quickly becomes worse when the average number of trucks approaches zero. Hence, the performance of the linear model is not as bad as the wMAPE implies in some cases. However, how good it then really is, is hard to say.

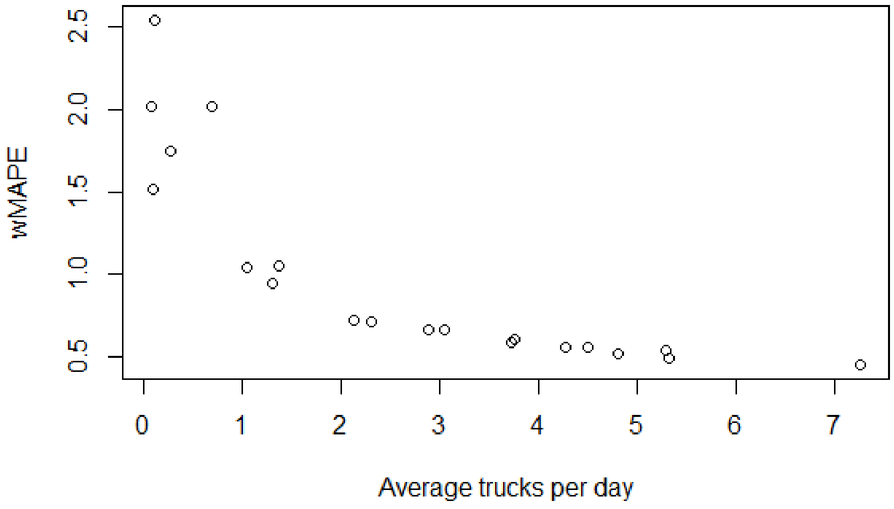


Figure 4.1: Scatter plot of the wMAPE of the linear model versus the average number of trucks used for small incidents per fire station.

Overall performance per day

Finally, we also look at the total daily number of trucks used for small incidents (over all fire stations and type-clusters). This enables us to compare all models through one value. Since we have no day in the test set on which there were no incidents at all in the whole region Amsterdam-Amstelland, we can compute both the MAPE and the wMAPE. This yields

$$\text{MAPE(LM)} = 0.1886 \quad \text{and} \quad \text{wMAPE(LM)} = 0.1924. \tag{5}$$

Since the wMAPE is higher, we can conclude that the LM is not very good at predicting relatively busy days (compared to predicting average days). However, the fire brigade is of course more interested in when they have busy days. They are prepared for average days anyway. We will therefore try to find another model that is better in predicting those busy days.

Conclusion 4.3 *Most weather variables which we expected to have predictive power for particular type-clusters confirmed our expectation, but not all of them. In judging the performance, it is important to account for the predictability of certain type-clusters or fire stations by looking at the CoV and the daily average number of trucks used.*

4.2.2 *Generalized linear model*

The idea for a generalized linear model (GLM) arises from an observation from Table 3.1. Here, we see that the largest outlier neither has the highest windspeed nor the most rainfall among those five outliers. However, the *combination* of wind and rainfall yet causes this day to be the largest outlier. It may therefore be a good idea to include also those cross-effects in our model, i.e.

$$\begin{aligned} x_{t,d} = \beta_0 &+ \beta_1 \cdot d \\ &+ \beta_2 \cdot \text{windspeed}_d \\ &+ \beta_3 \cdot \text{temperature}_d \\ &+ \beta_4 \cdot \text{rainfall}_d \\ &+ \beta_5 \cdot \text{visibility}_d \\ &+ \beta_6 \cdot \text{windspeed}_d \cdot \text{temperature}_d \\ &+ \beta_7 \cdot \text{windspeed}_d \cdot \text{rainfall}_d \\ &+ \beta_8 \cdot \text{windspeed}_d \cdot \text{visibility}_d \\ &+ \beta_9 \cdot \text{temperature}_d \cdot \text{rainfall}_d \\ &+ \beta_{10} \cdot \text{temperature}_d \cdot \text{visibility}_d \\ &+ \beta_{11} \cdot \text{rainfall}_d \cdot \text{visibility}_d \\ &+ \epsilon_{t,d}. \end{aligned} \tag{6}$$

Here, $\epsilon_{t,d}$ is again a residual term with zero expectation and some finite variance. Note that this is not a GLM as one may know from the literature. The only feature that causes it to be generalized is that it now also handles the cross-term relations between the weather variables. We could have called it an *expanded* linear model as well.

Model estimation

Note that this model is an expanded version of the linear model in Equation 4, so it should be at least as good. The question is how much value it adds to the linear model. It will therefore be interesting to investigate the estimated parameters for the cross-term variables, as well as to compare the performance of both models. Therefore, we first look at the predictive power of all GLM variables in Table 4.4 (the exact parameter values are given in the appendix in Table B.5). Compared to that of LM in Table 4.3, we observe that in general the single weather variables have lost some importance in favour of cross-term variables they partition in. Type-cluster 1 is an excellent example for this. Here, temperature had some predictive power in the LM, but now it turns out that it is mainly the *combination* with the amount of rainfall that matters. In addition, also windspeed and rainfall turn out to be less predictive on their own then the LM indicated. It is really their cross-term effect that is important. Looking at the average column on the right, we see that also the intercept has lost some importance. Apparently, a bigger part of reality can be modelled by the weather after adding some cross-term variables. Of all weather variables, it is even the case that two cross-term variables have most predictive power (on average).

Performance

Noting the influence of the cross-term variables, we expect that the performance of the GLM is better than that of the LM. However, Table B.8 shows that this is only the case for type-clusters 1 and 9. This is not surprising in the sense that these are the only type-clusters that have one or more cross-term weather variables with very significant predictive power (scale 1 or 2). It is disappointing though that it does not make any difference for the other type-clusters. Also if we judge it per fire station it hardly makes any difference (see Table B.9). Fortunately, if we compute the results for the totals per day, we get

$$\text{MAPE}(\text{GLM}) = 0.1865 \quad \text{and} \quad \text{wMAPE}(\text{GLM}) = 0.1880. \tag{7}$$

Still, the wMAPE is somewhat higher than the MAPE, but compared to their equivalents of the LM (see Equation 5) they are slightly better (about 2%).

Table 4.4: Significance on a 1-5 scale of the parameters of the estimated generalized linear models. Remember that a different model is estimated for each type-cluster. The scale is based on the p-value of two-sided *t*-tests.

	Type cluster									
Variable	1	2	3	4	5	6	7	8	9	Avg
Intercept	1	2	1	1	1	5	1	2	3	1.89
Trend	1	5	1	4	3	5	5	5	5	3.78
Windspeed	3	5	5	5	5	5	5	5	1	4.33
Temperature	5	5	5	3	2	5	5	5	4	4.33
Rainfall	5	3	5	5	5	5	5	5	1	4.33
Visibility	5	5	4	5	5	5	5	5	5	4.89
Wind*Temp.	5	5	5	5	5	5	5	5	5	5.00
Wind*Rain	3	3	5	5	5	5	5	5	1	4.11
Wind*Visib.	5	3	5	4	5	5	5	5	5	4.67
Temp.*Rain	2	3	5	5	5	5	5	5	1	4.00
Temp.*Visib.	5	5	5	5	5	5	5	5	5	5.00
Rain*Visib.	5	5	5	5	5	5	5	3	5	4.78
Scaling: 1: p<0.001, 2: p<0.01, 3: p<0.05, 4: p<0.1, 5: p<1										

Conclusion 4.4 *GLM performs better than LM, because it turns out that sometimes it is specifically the combination between variables that is important, especially rain in combination with wind and temperature.*

4.2.3 *Random forest*

The random forest (RF) algorithm is a machine learning algorithm that can be used for both classification and regression tasks. Compared to LM and GLM it has a large computation time, but RF is often used in practice since it generally has great performance. It will therefore be worth a try to implement this algorithm for our regression problem.

How the algorithm works

As input, the algorithm needs a $T \times (K+1)$-matrix with K explanatory variables (can be both numeric and categorical) and one observation variable (in this case $x_{t,d}$), all of sample size T. In the first iteration of the algorithm, a sample of size T (in our case) is drawn with replacement from the input matrix. On this sample, a *decision tree* (DT) algorithm is executed. DT starts off with a root node containing the complete sample. Then this sample is cut into two subsamples in such a way that the reduction in total (weighted) standard deviation is maximal. For us, this means that the days within each subsample have more similar characteristics between each other than with respect to those in the other subsample. For instance, one subsample may include all days from the sample in which the total rainfall was more than 5mm, so that $x_{t,d}$ is typically larger/smaller in this subsample than in the other one. This splitting into subsamples continues until no splitting is possible anymore. A criterion here is that each so-called leaf node must contain a subsample of size at least 5 (in our case). If now a new sample comes in, we can check to which leaf node it belongs and then predict its value $x_{t,d}$ by taking the average of the observations in this leaf node. This procedure is repeated N times, yielding N decision trees (that's why it is called a forest). When a new sample comes in, we can take all N predictions for this sample (one from each decision tree) and average these to get the final prediction. We will see later how big we must take N as to balance the computation time opposed to the quality of the prediction.

Implementation

Since computation time plays a much bigger role here than with the LM and GLM, we will implement RF first with only the single weather variables (as in LM). Then, we will implement it again for both the single and cross-term weather variables (as in GLM). The first question of interest here is how much longer the computation time is for the second implementation and how much better the performance becomes. Secondly, we want to compare of course the performance of both implementations to that of the LM and GLM.

Settings

INCLUDING CROSS-TERM VARIABLES We will run the algorithm under different settings in order to determine which settings are most suitable. The first question is whether we should include the cross-term weather variables in the input matrix of the algorithm. Unlike the linear models, RF should be able to capture some of the cross-term relationships without including those specific variables. Running the algorithm both with and without the cross-term variables yields that the running time differs by almost a factor 2. Moreover, the difference in MSE is indeed negligible, so we can safely exclude the cross-term variables from the input matrix.

NUMBER OF DECISION TREES Second, we want to determine what number of trees (N) to use. Figure 4.5 shows the decrease in Mean Squared Error (MSE) when we execute the algorithm for the first type-cluster. Initially, we used $N = 2723$, which equals the number of days in the training set. Apparently, the MSE reduces fast in the beginning, but this reduction stabilizes quickly as well. On the other hand, after some experimenting it turns out that the running time of RF is approximately linear in the number of trees it computes, which is not surprising regarding the way the algorithm works. This means that RF runs more than 5 times faster when we limit the number of trees to 500, while we then lose only about 0.4% in MSE. In this case the additional running time is not worth it, so we will use $N = 500$ in our implementation. Under these settings, the RF algorithm runs in less than ten seconds per type-cluster, so this is very acceptable. In future implementations, using (say) $N = 100$ trees would be fine as well, but much less is not recommended.

Table 4.5: Table of the MSE reduction in the random forest algorithm for the third type-cluster (the pattern is comparable for all type-clusters, only the scale differs). The reduction goes along with computing more decision trees, but after some time, adding more trees is not very rewarding anymore.

# Trees	1	2	5	10	20	50	100	250	500	2723
MSE	15.33	14.22	12.35	10.69	9.49	8.94	8.68	8.48	8.45	8.42

VARIABLE SELECTION The only variables we use are *exactly* four weather variables. An intercept is not needed here and a trend variable is not useful in this algorithm. Note that for the linear models, we can define a trend variable for the training set by just taking a vector from 1 to N_{training} and for the test set by a vector from $N_{\text{training}} + 1$ to $N_{\text{training}} + N_{\text{test}}$. However, the decision trees in the RF algorithm are build just on the training set. So these will treat all trend values from the test set the same, since they are all bigger than the maximum trend value in the training set. Hence, adding a trend variable will rather have a negative than a positive effect on the forecasts, so we better not do this. Furthermore, since RF is able to capture non-linear relationships and correlation is a linear measure, we cannot say that very low correlations in Table 4.2 indicate that we can leave out that particular weather variable from our model. Therefore, we do always choose one variable from each of all four categories. The reason to not include all variables is to keep the running time low. Now, before discussing the results, realise that the RF algorithm involves generating *random* samples, so the results may slightly differ among consecutive runs (even when using the same settings). The following results are taken from just one of these runs for each type-cluster.

Results

We will first dive into the predictive power of the variables again. Different from the previous models, the RF algorithm does not estimate a parameter for each variable. We therefore have to find another measure for the importance of each variable. We will consider two options, which we will call 'RSS-ranking' and 'root-ranking'.

RSS-RANKING Remember that in each decision node, the algorithm splits the remaining sample based on a decision rule on the variable that reduces the standard deviation most. In other words, it tries to improve the fit of the model to the training data as much as possible, i.e. the biggest decrease in *residual sum of squares* (RSS) between the fitted model and the observation data in the training set. Hence, we can measure the importance of a variable based on the total decrease in RSS from splitting on this variable. Therefore, Table B.6 in the appendix gives these numbers for all variables, but what is most interesting for us is the ranking of the variables based on these numbers as in Table 4.6. As in the previous models, visibility is often the least important variable. However, the biggest difference is that in this case the temperature is remarkably important.

Table 4.6: The importance of variables according to a ranking based on the total decrease in RSS from splitting on each of these variables. Here, 1 indicates the most important variable for each type-cluster.

	Type-cluster									
Variable	1	2	3	4	5	6	7	8	9	Avg
Windspeed	4	2	4	1	3	2	2	4	1	2.56
Temperature	1	1	1	3	2	1	1	1	3	1.56
Rainfall	3	4	3	2	1	4	4	3	2	2.89
Visibility	2	3	2	4	4	3	3	2	4	3.00

ROOT-RANKING For the second importance measure, we will look at the root node of each of the 500 decision trees created by the algorithm. In Table B.7 in the appendix, we show the number of times each variable has been the decision variable in the root node of a decision tree. A ranking based on these counts is given in Table 4.7. This ranking surely resembles the previous one on a bigger scale. However, if we look at both rankings in detail, there are some obvious differences. First, the root-ranking seems to resemble better the relative importance of the variables from Table 4.3. Second, the RSS-ranking gives a more subtle ranking in the sense that it just ranks from one to five. The root-ranking assigns a five to multiple variables when they both are never used as decision variable in the root node. Moreover, the root-ranking attempts to find the most important variable in every decision tree, but it does not explicitly regard the ranking below the most important one. So in short, the RSS-ranking theoretically seems to be more subtle, but in practice the root-ranking seems to make more sense *in this case*.

Table 4.7: The importance of variables according to a ranking based on the frequency of being the decision variable in the root node of a decision tree in the random forest algorithm. Here, 1 indicates the most important variable for each type-cluster.

	Type-cluster									
Variable	1	2	3	4	5	6	7	8	9	Avg
Windspeed	4	4	1	4	4	1	1	2	1	2.44
Temperature	4	1	2	2	2	2	2	1	2	2.00
Rainfall	1	2	4	3	1	4	4	3	4	2.89
Visibility	4	3	3	1	3	3	3	4	4	3.11

Performance

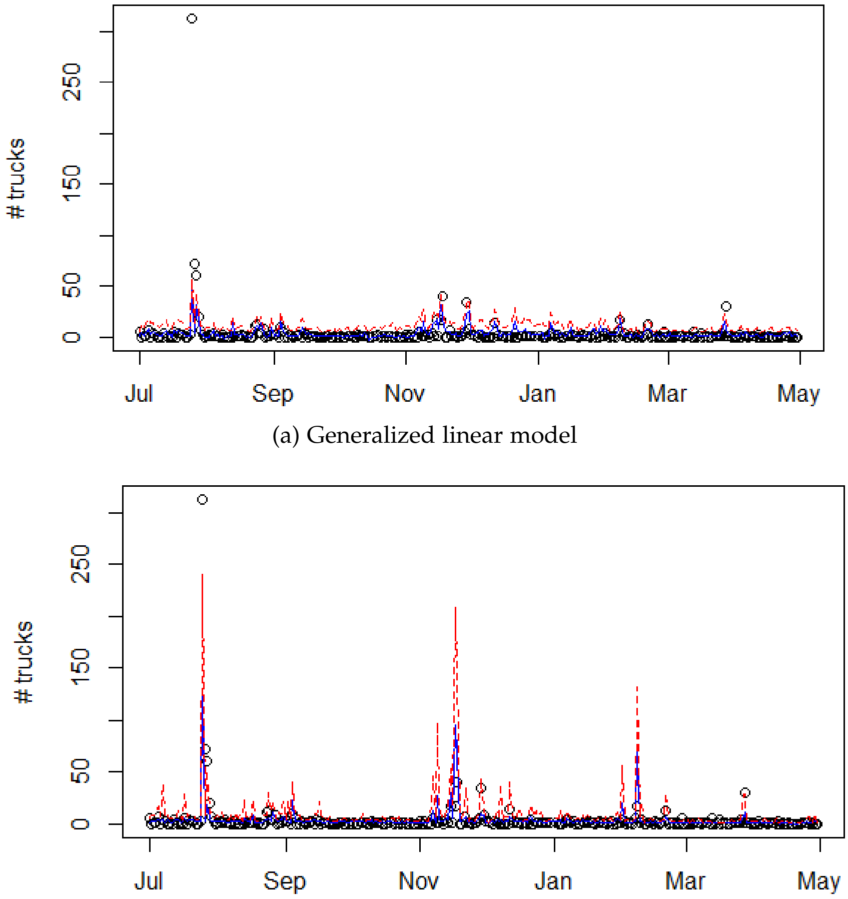
Now it is time to compare the results of RF to the previous models. Therefore, we look again at Table B.8 and see that in general, RF gives the worst results. Is all the effort we put in this model for nothing? Maybe not, because RF has the best wMAPE for type-cluster 9, which may be an indication that this algorithm is better in predicting busy days. This is confirmed by the plot of the predictions for type-cluster 9 of both GLM and RF in Figure 4.2. Obviously, the RF algorithm recognizes much better than GLM when the weather conditions are risky and likely to cause many incidents to happen (look at the big outlier in July).

Finally, the results for the totals per day are

$$\text{MAPE(RF)} = 0.2006 \quad \text{and} \quad \text{wMAPE(RF)} = 0.2019. \tag{8}$$

Although the results are slightly worse than for the previous models, it is hopeful that the results for type-cluster 9 are better. We will try to use this fact in the next section.

Conclusion 4.5 *Overall, RF performs worse than both linear models, but it is better in predicting busy days. Furthermore, it turns out that temperature has very specific (non-linear) predictive power, which could not be captured by the linear models.*



(a) Generalized linear model

(b) Random forest

Figure 4.2: Forecasts (in blue) of the number of trucks used for small incidents of type-cluster 9, including the upper bound of its 95%-prediction interval (in red).

4.2.4 *Ensemble averaging*

From the previous sections, we can conclude that GLM gives the best results when we look at the totals per day, but it is worse in predicting busy days than RF. If we can combine both models in such a way that we capture the good features from both worlds, than this may improve our forecasts. We will try to do this by applying a form of so-called *ensemble averaging* (EA). In our case, we will take a weighted average of the forecasts of RF and GLM, i.e.

$$\mathrm{EA} = \gamma \cdot \mathrm{RF} + (1 - \gamma) \cdot \mathrm{GLM}, \tag{9}$$

for some constant $\gamma \in [0, 1]$.

Performance

We have to find out which value of γ to use in order to get the best results. Since GLM initially gives the best results and we only need RF to be able to predict the busy days a bit better, we may expect that we have to put more weight on GLM, i.e. that $\gamma < 0.5$. Table 4.8 confirms this expectation. Both the MAPE and the wMAPE take their minimum in $\gamma^* = 0.2$ (which is hence better than GLM individually; compare with $\gamma = 0$). Using this value for γ, we can also compare the results if we look at the predictions per type-cluster (Table B.8) or per fire station (Table B.9). Note that its performance is not always better than for the other models. However, we still prefer this EA method, because it is a good balance between accurately predicting average days and being able to foresee busy days. Therefore, we will also use the predictions from this model in the following sections.

Table 4.8: Results of ensemble averaging when looking at totals per day, using $\text{EA} = \gamma \cdot \text{RF} + (1 - \gamma) \cdot \text{GLM}$

γ	MAPE	wMAPE
0	0.1865	0.1880
0.1	0.1854	0.1863
0.2	0.1853	0.1860
0.3	0.1858	0.1865
0.4	0.1868	0.1876
0.5	0.1880	0.1889
0.6	0.1895	0.1905
0.7	0.1915	0.1926
0.8	0.1942	0.1954
0.9	0.1973	0.1985
1	0.2006	0.2019

Conclusion 4.6 *As our final model for the small incidents, we will use ensemble averaging with $\gamma = 0.2$, i.e. $EA = 0.2 \cdot RF + 0.8 \cdot GLM$. Using this model, we do better in predicting both regular and busy days than any of the considered models could do on its own.*

4.3 FIREWORK INCIDENTS

Now that we have forecasts for the number of trucks used for small incidents per day, we can finally have a closer look on the New Year's days (recall Section 3.1). Note that the models we used in the previous sections are all based on the data without both the big incidents and all incidents on every December 31 and January 1. If we now use these models to make a forecast for these New Year's days, then this forecast is based only on 'normal' days in that time of the year, the weekday and the weather conditions. However, we already concluded that these days are not 'normal' because the use of fireworks causes many extra incidents to happen. What we will do now is make a forecast for all New Year's days based on the EA method and subtract this from the real number of trucks used for small incidents on these days. We will make the assumption that the result approximates the number of firework incidents, so that we can analyse them separately.

Results

First, we look at the correlations with the weather variables in Table 4.9. Apparently, people use more firework on cold New Year's days with not too much wind and rain, which sounds plausible. In any case, more incidents seem to happen under these weather conditions.

Table 4.9: Correlation between weather variables and firework incidents.

Variable	Correlation	p-value
Windspeed (FG)	-0.679	0.003
Temperature (TG)	-0.667	0.003
Rainfall (DR)	-0.575	0.016
Visibility (VVN)	-0.407	0.104

Note though that we only have 17 New Year's days in our dataset. Moreover, recall from Figure 3.2 that in 2014/2015 the allowed time interval for using fireworks has been changed. Therefore, we only have two years in our datasets that are completely representative for future New Year's days. Hence, putting much effort in modelling these firework incidents is not really worth it for now, so we will only implement a simple linear model with just an intercept and four weather variables. The results of estimating the model on all New Year's days are given in Table 4.10. None of the weather variables have significant predictive power, which may also just be because of the lack of data. If we estimate the model just on the first twelve New Year's days and leave the last five for testing, then we get a MAPE of 0.349. But again, this does not say much since we have too few days to make a good forecast. However, this is the only way we can deal with the different patterns on New Year's days, so we will model it like this anyway.

Table 4.10: Parameter estimates of a linear model for the firework incidents, including p-values of two-sided *t*-tests to test the null hypothesis that the true parameter equals zero.

Variable	Estimate	p-value
Intercept	219.158	0.000
Windspeed (FG)	-0.607	0.352
Temperature (TG)	-0.565	0.198
Rainfall (DR)	0.041	0.941
Visibility (VVN)	-0.405	0.519

Conclusion 4.7 *It seems that more people use firework on cold New Year's days with not too much wind and rain, but none of these weather variables have significant power in predicting (what we assumed to be) firework incidents. Because of the little data on New Year's days, we will just use a simpel LM for predicting firework incidents.*

4.4 NEEDED CAPACITY FIRE STATIONS

Now that our forecasts are complete, it is interesting to extract from this the capacity we expect each fire station to need each day. For this, we want to have some certainty that the capacity is satisfying for that day. Different from a confidence interval, which only measures the uncertainty of the forecast, a prediction interval includes in addition the variability of the number of incidents in real life. We can therefore use the upper bound of the prediction interval to ensure that the predicted capacity will be satisfactory with, for instance, 95% certainty. But how can we compute such intervals for GLM and RF?

Prediction interval for GLM

The standard function in *R* to make a prediction interval for the (multiple) linear regression model

$$y = X^{T}\beta + \epsilon$$

assumes that all elements of $\epsilon = (\epsilon_1, \dots, \epsilon_n)$ are independent and follow an identical normal distribution, i.e. $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ for all t independently. Then, the $100(1-\alpha)$%-prediction interval for a future observation y_0 can be computed as

$$\widehat{y}_0 \pm t_{n-k}^{(1-\alpha/2)} \sqrt{\widehat{\mathbb{V}\mathrm{ar}}(X^{T}\beta + \epsilon)} = \widehat{y}_0 \pm t_{n-k}^{(1-\alpha/2)} \widehat{\sigma} \sqrt{x_0^{T}(X^{T}X)^{-1}x_0 + 1},$$

(see [5]) where

- $\widehat{y}_0$ is the predicted value for y_0;
- $t_{n-k}^{(1-\alpha/2)}$ is the $(1-\alpha/2)$-quantile of the t-distribution with $n-k$ degrees of freedom;
- n is the number of samples in the training set;
- k is the number of variables in the model;
- and $\mathbb{V}\mathrm{ar}(X^T\beta+\epsilon)$ includes both the uncertainty of the prediction ($\mathbb{V}\mathrm{ar}(X^T\beta)$, i.e. the confidence interval part) and the variability of the observations ($\mathbb{V}\mathrm{ar}(\epsilon)$).

We can write the prediction interval in a compact way like this (using ±), because the t_n-distribution is symmetric around zero. A more general way of denoting the same interval is

$$\left[\widehat{y}_0+t_{n-k}^{(\alpha/2)}\widehat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0+1},\widehat{y}_0+t_{n-k}^{(1-\alpha/2)}\widehat{\sigma}\sqrt{x_0^T(X^TX)^{-1}x_0+1}\right].$$

Now, recall that for $n\to\infty$ the t_n-distribution converges to a standard normal distribution. Hence, σt_n then converges to a $\mathcal{N}(0,\sigma^2)$-distribution, which is the distribution of ϵ_t according to the standard assumption in *R*. But in our case, this assumption does not hold (recall Section 4.2.1). Instead of trying to fit another distribution for ϵ_t, we will just assume that the residuals in future observations will have the same distribution as the residuals of our fitted model on the training set. Then, we can use the quantiles of these residuals (denoted by q_α) and compute the prediction interval of GLM as

$$\left[\widehat{y}_0+q_{\alpha/2}\sqrt{x_0^T(X^TX)^{-1}x_0+1},\widehat{y}_0+q_{1-\alpha/2}\sqrt{x_0^T(X^TX)^{-1}x_0+1}\right].$$

Prediction interval for RF

Remember that the RF algorithm computes N decision trees, which all yield one prediction for each future observation. The variability of these N individual predictions captures the uncertainty of the final prediction (the average of the individuals). In order to capture the variability of the observations, we need again our assumption on the residuals. In this case, we will use this by adding to each of the N individual predictions a random value, drawn from the empirical distribution of the residuals in the training set. Then, the resulting N values include all the variation we need. Their $(\alpha/2)$- and $(1-\alpha/2)$-quantiles together directly form the desired prediction interval.

Results

Now, we have to combine all previous results. If we want to be sure that the capacity suffices with $100(1-\alpha)\%$ certainty, we need (of course per day and per fire station) the following ingredients:

A. The upper bound of the $100(1-2\alpha)\%$-prediction interval of the EA forecast for small non-firework incidents.

B. The upper bound of the $100(1-2\alpha)\%$-prediction interval of the LM forecast for small firework incidents.

C. The (decreasing) $(1-\alpha)$-quantiles for the big incidents.

Then, we can determine the needed capacity by

$$(A+B+C)\cdot\frac{\text{maximum hour ratio}}{24},$$

where the maximum hour ratio is approximately 1.21 (recall Figure 3.6c). Table 4.11 gives the needed capacity for each fire station.

Table 4.11: The capacity needed per day for each fire station if it wants to have the given certainty that the capacity suffices that day. In our test set (2015/07–2016/04) no fire station ever needed a capacity of more than two trucks.

	Avg cap. needed			% of days 2 needed			Available
Fire station	90%	95%	99%	90%	95%	99%	cap. 1?
Aalsmeer	0.14	0.17	0.27	0.0%	0.0%	0.0%	No
Amstelveen	0.44	0.53	0.80	0.0%	0.3%	3.3%	No
Anton	0.40	0.48	0.73	0.0%	0.0%	0.3%	No
Diemen	0.12	0.15	0.25	0.0%	0.0%	0.0%	No
Dirk	0.34	0.41	0.64	0.0%	0.0%	0.7%	No
Driemond	0.04	0.05	0.10	0.0%	0.0%	0.0%	Yes
Duivendrecht	0.17	0.20	0.30	0.0%	0.0%	0.0%	No
Hendrik	0.59	0.71	1.07	0.7%	1.7%	67.7%	No
IJsbrand	0.19	0.24	0.38	0.0%	0.0%	0.0%	Yes
Landelijk Noord	0.04	0.06	0.11	0.0%	0.0%	0.0%	Yes
Nico	0.35	0.42	0.64	0.0%	0.0%	0.3%	No
Osdorp	0.42	0.51	0.77	0.0%	0.0%	1.0%	No
Ouderkerk a/d Amstel	0.06	0.08	0.13	0.0%	0.0%	0.0%	Yes
Pieter	0.41	0.50	0.75	0.0%	0.0%	1.7%	Yes
Teunis	0.28	0.34	0.53	0.0%	0.0%	0.0%	No
Uithoorn	0.12	0.15	0.25	0.0%	0.0%	0.0%	No
Victor	0.28	0.34	0.51	0.0%	0.0%	0.0%	No
Willem	0.30	0.36	0.55	0.0%	0.0%	0.0%	No
Zebra	0.23	0.28	0.44	0.0%	0.0%	0.0%	Yes

From this we can conclude that, on an average day, (almost) all fire stations only need a capacity of one truck. Only if we want to be 99% sure that the capacity suffices, we need a capacity of two trucks at station 'Hendrik' on an average day. More specifically, with this certainty

Hendrik needs a capacity of two trucks on more than two thirds of the days. Even when only a certainty of 90% is required, then there are some days on which Hendrik needs two trucks. If the required certainty is increased to 95%, then Amstelveen joins the club of stations who need a capacity of two on some days. If we finally increase the required certainty to 99%, the only station that gets in trouble is Pieter. Then, on 1.7% of the days the available capacity of Pieter does not meet the desired certainty of being sufficient (see in red). However, recall that the demand is slowly decreasing. Moreover, even on these 'critical' days the capacity does ensure to be sufficient with *at least* 95% certainty, so there is no reason to panic. Note that we can say 'at least', because in reality we need to round up to ensure the desired certainty for an integer capacity. However, since the needed capacity is often (if not always) just 1 or 2, this rounding up may cause the achieved certainty to be much higher than initially asked for.

Conclusion 4.8 *If we want to be 99% certain that the capacity suffices, then only fire station 'Pieter' is not able to meet this criterion on 1.7% of the days. Even on these days, Pieter does have at least 95% certainty of having sufficient capacity. So we can conclude that no fire station in Amsterdam-Amstelland has to expand its capacity.*

5 FORECAST PER REGION

In this final chapter, we will briefly discuss a second approach of modelling. In the previous chapter, we computed a different model for each type-cluster and then divided the total prediction among the fire stations. This is the easiest way to obtain a forecast per fire station. However, this is not the most intuitive way of predicting incidents, because incidents happen at a certain place *in a certain region*, and not at a certain place where (at that moment) a truck of fire station X was closest by. It would therefore be more logical to make a forecast per region and then allocate the demand among all fire stations. A benefit of this approach is that we can more accurately use the characteristics of regions. For example, a region with many (big/old) trees may be risky on stormy days. Or, a region with a high density of houses may be more prone to inside fires. It is outside the scope of this paper to fully implement this approach in detail, but in this chapter we will briefly discuss how this can be achieved.

5.1 MODELLING AND RESULTS

In this research, we first created a model for each type-cluster and then divided the prediction over all fire stations. We can try to capture some of the region characteristics by first dividing our prediction per type-cluster over all regions according to the percentages in Table B.10. When we use the LM of Section 4.2.1, then the results per region are as illustrated in Figure 5.1 (exact results given in Table B.11). In this plot every point represents a region, so we can again conclude that the less incidents happen, the harder it is to make a good forecast in terms of wMAPE. Comparing these results to those in Table B.9, then we observe that the predictions per region are better (except for 'Buiten regio'), but that this is only due to the size of the groups we created. For the totals per day, we get

$$\text{MAPE}(\text{LM2}) = 0.1887 \quad \text{and} \quad \text{wMAPE}(\text{LM2}) = 0.1919,$$

which is comparable to the previous results of LM (see Equation 5). This is not surprising since we used exactly the same models for the type-clusters first. The biggest challenge for future research is to come up with a model that can be applied separately to each region, while it still accounts for all different incident types. If such a model can be made, then we may expect it to yield better results than achieved in this research.

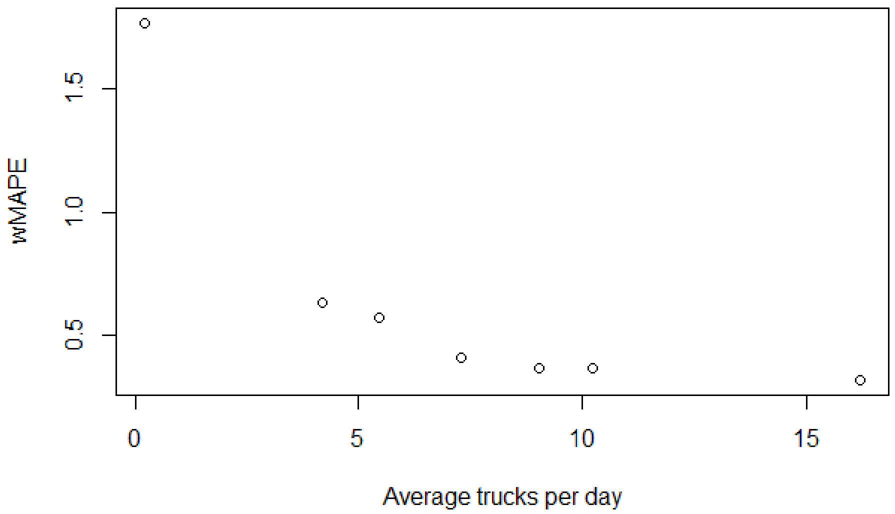


Figure 5.1: Scatter plot of the wMAPE of the linear model versus the average number of trucks used for small incidents per region.

5.2 ALLOCATION OVER FIRE STATIONS

Having a prediction per region, the remaining question is how we can divide this prediction over all fire stations. After all, this is what is most interesting from the fire brigade's point of view. The allocation rule must of course be based on the average response time of fire station x when an incident happens in region y. Because the response times are not known, we will use the average distances $d(x,y)$ here instead. The most simple allocation rule may be

$$s(x,y) = \frac{d(x,y)^{-1}}{\sum_i d(i,y)^{-1}}, \tag{10}$$

where $s(x,y)$ denotes the share of fire station x in the prediction of region y. We need inverse distances to achieve that shorter distances result in higher shares. The numerator is just a normalization to make all shares for each region sum up to one. Unfortunately, this rule yields an allocation that it very different from the allocation based on the data (compare Tables B.13 and B.14). Instead of searching for another allocation rule, we will just use the percentages as given in Table B.14. Then, the results per fire station are as given in Table B.12. These show that the results are slightly better on average, although the difference is small. Nevertheless, it gives good hope that this alternative approach may yield better results when it is executed in a more advanced way.

Conclusion 5.1 *In future research, improvements can be aimed for by trying to capture region characteristics while still accounting for all different incident types.*

6 CONCLUSIONS AND RECOMMENDATIONS

6.1 CONCLUSIONS

During this research, we tried to find an answer to the question:

Can we make a good forecast on the number of incidents that each fire station in Amsterdam-Amstelland has to handle?

Here, special interest went to the influence of several weather conditions and to the issue of dealing with the low number of incidents.

Unless the main question is somewhat subjective, we can be satisfied with the forecasts we created for the small incidents using ensemble averaging (EA). Also the result that the big incidents can be modelled by an inhomogeneous Poisson process is quite nice. We found that the incident types are very useful in predicting the small incidents, but since some of them do not occur frequently enough, we had to split the relevant types into nine clusters. Here, we ensured that each type-cluster contained at least one incident per day. Because of a lack of data, we could not make a separate model for each type-cluster/fire station combination. This issue was partly solved by only modelling each type-cluster separately and then dividing the predictions over the fire stations. Concerning the weather, (the combination of) rain and wind on average had most influence in the linear models and temperature appeared to contain mostly non-linear relations with the number of incidents. As expected beforehand, the visibility had the least predictive power among those four weather variables.

In the end, the linear model with cross-term effects of the weather variables (GLM) turned out to yield the best results in terms of weighted mean absolute percentage error (wMAPE). On the other hand, the random forest algorithm (RF) showed to be better in predicting busy days, which is of course more interesting for the fire brigade. Taking an optimized weighted average between these two models ($\text{EA} = 0.2 \cdot \text{RF} + 0.8 \cdot \text{GLM}$) finally seemed to capture the best of both worlds ($\text{wMAPE} = 0.1860$ for total number of small trucks per day). Using the prediction intervals of this model, we finally concluded that the available capacity of fire trucks suffices for all fire stations.

6.2 RECOMMENDATIONS

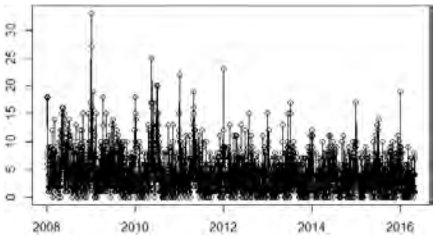
TYPE-CLUSTERING The type-clustering we made in this research was based on correlations with the weather variables. This turned out not too bad, but there are several reasons why this clustering may not be optimal. First, it is only based on *linear* relations to the concerned weather variables while for example temperature showed to have mainly influence in a non-linear way. Second, we saw for example that inside fires occur more frequently during weekends and that certain type of incidents typically require multiple trucks. We can also include these kind of characteristics in determining a clustering. Furthermore, the weekday corrections were done per type-cluster here, but it may be that there are types within one cluster that have opposing weekday factors. It may therefore be wise to use different clusterings, one for the corrections and one for the models using the weather conditions. These are interesting topics to investigate in future research.

RANDOM FOREST Another improvement that can be made is in the RF algorithm. We discussed why we cannot include a trend *in* the algorithm, but there is no reason why we cannot apply a correction for the trend before implementing the algorithm. For time reasons, we did not implement this idea in this research, but it will almost certainly lead to better results (although the difference may be small).

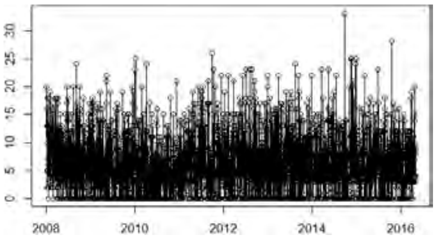
FORECAST PER REGION Finally, improvements can be made by investigating the opportunities raised by the approach in Chapter 5. Already for my simple implementation, the results of first looking at a prediction per region instead of per fire station showed that this approach has a great potential.

A FIGURES

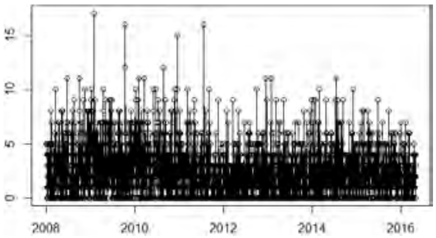
TYPE CHARACTERISTICS



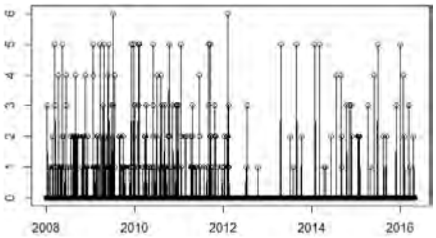
(a) Buitenbrand



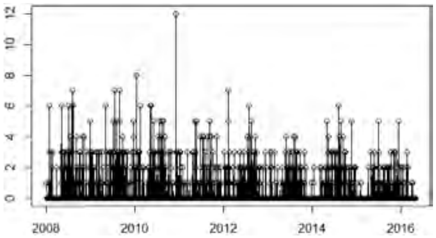
(b) Binnenbrand



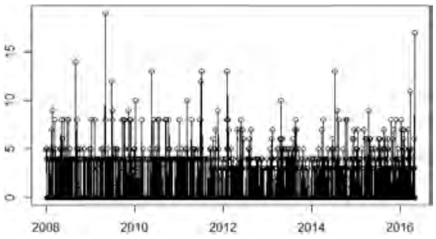
(c) Brandgerucht / nacontrole



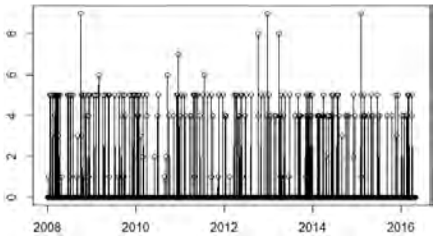
(d) Hulpverlening water algemeen



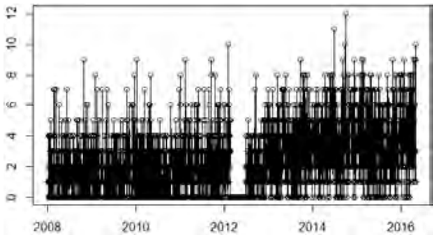
(e) Dier te water



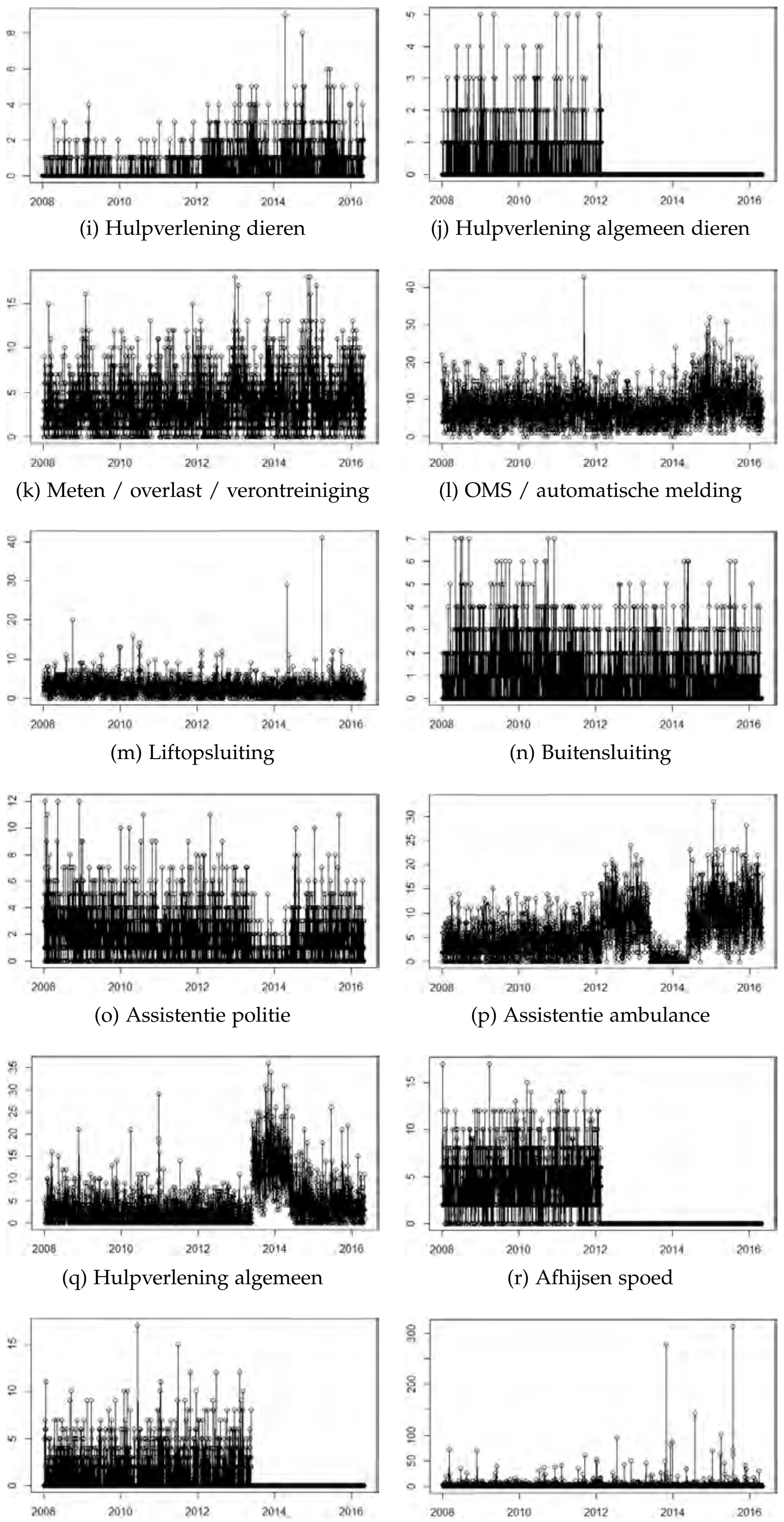
(f) Persoon te water



(g) Voertuig te water



(h) Reanimeren



(i) Hulpverlening dieren

(j) Hulpverlening algemeen dieren

(k) Meten / overlast / verontreiniging

(l) OMS / automatische melding

(m) Liftopsluiting

(n) Buitensluiting

(o) Assistentie politie

(p) Assistentie ambulance

(q) Hulpverlening algemeen

(r) Afhijsen spoed

(s) Beknelling / bevrijding

(t) Storm en waterschade

Figure A.1: For each type with more than 100 small incidents in total a plot of the number of trucks used for small incidents per day.

EXTRA FIGURES FOR SEASONAL PATTERNS

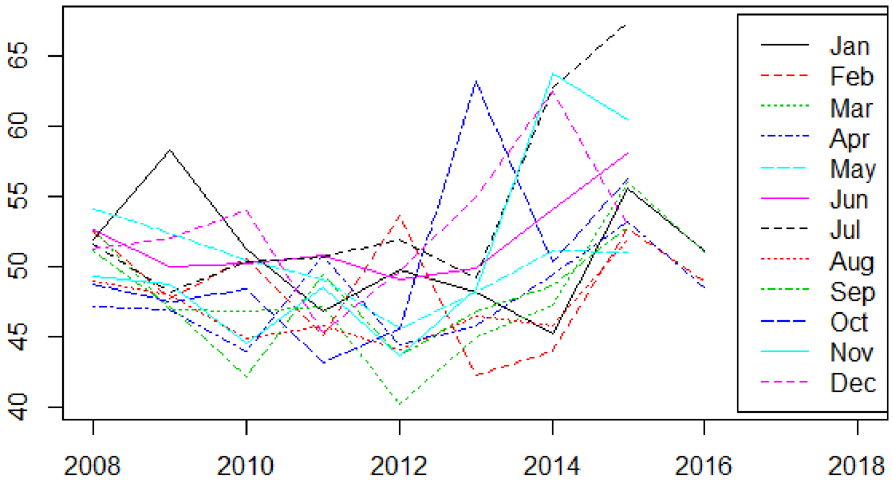
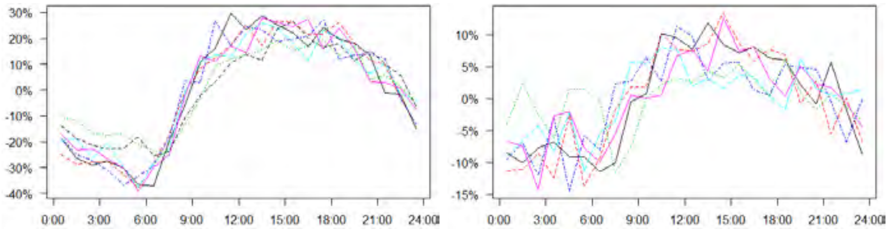


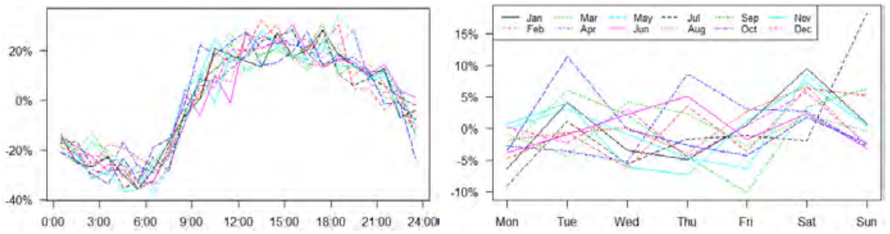
Figure A.2: For every month a plot of the yearly development of the number of trucks used for small incidents per day.



(a) Per weekday.

(b) Per cluster.

Figure A.3: Day patterns per weekday and per cluster. There are no significant differences within each plot.



(a) Day pattern.

(b) Week pattern.

Figure A.4: Day and week pattern throughout the year. The day pattern does not really change, but the week pattern does seem to be slightly changing.

B TABLES

RELEVANCE INCIDENT TYPES

Table B.1: The number of small incidents (excluding New Year's days) per incident type, together with their first and last occurrence. The types in the lower block either occur only rarely or do not exist any more. Note that there are 3026 non-New Year's days in total, so even some incident types in the upper block do not even occur on a daily basis.

	Incident type	Freq.	First date	Last date
1	Assistentie ambulance	9552	01/01/2008	29/04/2016
2	Assistentie politie	4360	01/01/2008	28/04/2016
3	Binnenbrand	5407	01/01/2008	29/04/2016
4	Brandgerucht / nacontrole	5163	01/01/2008	29/04/2016
5	Buitenbrand	10459	01/01/2008	29/04/2016
6	Buitensluiting	2216	01/01/2008	09/04/2016
7	Dier te water	785	01/01/2008	01/04/2016
8	Hulpverlening algemeen	7092	01/01/2008	29/04/2016
9	Hulpverlening dieren	1048	19/01/2008	29/04/2016
10	Hulpverlening water algemeen	241	08/01/2008	19/04/2016
11	Liftopsluiting	7434	01/01/2008	29/04/2016
12	Meten / overlast / verontreiniging	7632	01/01/2008	29/04/2016
13	OMS / automatische melding	17266	01/01/2008	29/04/2016
14	Persoon te water	867	01/01/2008	26/04/2016
15	Reanimeren	5912	01/01/2008	29/04/2016
16	Storm en waterschade	6369	01/01/2008	26/04/2016
17	Voertuig te water	230	03/01/2008	31/03/2016
18	NVT	38	27/01/2008	01/07/2015
19	Herbezetting	25	28/03/2012	18/06/2015
20	Brandbare gassen	66	17/10/2012	16/04/2014
21	Regionale bijstand	12	14/07/2008	14/10/2013
22	Overige gevaarlijke stoffen	17	08/07/2008	06/09/2013
23	Beknelling / bevrijding	938	08/01/2008	23/05/2013
24	Afhijsen spoed	3262	01/01/2008	19/02/2012
25	Hulpverlening algemeen dieren	509	05/01/2008	19/02/2012
26	Interregionale bijstand	83	05/02/2008	16/02/2012
27	Brandbare vloeistoffen	2	09/09/2011	30/09/2011
28	Letsel eigen personeel	12	21/09/2008	22/05/2011
29	Buiten dienststelling	1	04/02/2011	04/02/2011

FIRE STATION/TYPE-CLUSTER COMBINATIONS

Table B.2: Average number of trucks needed for small incidents *per month* for each type-cluster separated per fire station.

Station \ Type-cluster	1	2	3	4	5	6	7	8	9
Aalsmeer	2.3	0.7	1.3	4.5	2.0	0.4	5.8	2.1	2.0
Amstelveen	7.4	5.4	4.8	35.7	6.1	3.9	22.4	11.8	5.4
Anton	8.6	7.4	4.7	35.9	10.3	3.4	18.5	4.7	5.0
Diemen	2.5	1.7	1.6	5.3	2.1	0.9	5.6	1.3	1.6
Dirk	5.4	4.4	5.7	14.5	8.4	2.5	19.7	4.3	5.4
Driemond	0.2	0.5	0.1	0.0	0.1	0.0	0.4	0.1	0.1
Duivendrecht	2.4	1.4	1.0	16.1	2.1	0.5	3.4	0.6	0.8
Hendrik	7.3	10.4	13.1	20.3	15.7	5.7	40.4	5.3	10.1
IJsbrand	8.9	2.8	3.0	6.6	4.1	1.9	10.3	3.3	2.7
Landelijk Noord	0.2	1.2	0.0	0.1	0.1	0.0	0.0	0.1	0.0
Nico	5.7	7.1	5.1	21.4	8.1	3.3	18.2	3.3	4.7
Osdorp	13.4	6.8	7.2	18.6	9.7	3.9	24.9	7.4	5.1
Ouderkerk a/d Amstel	0.8	0.7	0.2	0.4	0.5	0.1	1.5	0.3	0.3
Overig	0.3	0.2	0.1	0.6	0.2	0.0	0.1	2.9	0.1
Pieter	10.5	5.1	7.6	20.4	9.6	2.9	21.3	7.2	5.0
Teunis	6.3	4.8	5.6	12.6	7.3	2.2	13.3	3.4	3.6
Uithoorn	2.2	1.6	1.4	5.3	1.6	0.7	5.0	1.8	1.7
Victor	5.7	4.9	5.6	8.6	8.0	2.7	17.1	2.0	3.3
Willem	6.9	4.8	5.2	11.1	7.0	3.1	18.9	3.4	3.8
Zebra	7.5	5.2	2.8	9.1	4.9	2.3	11.9	3.5	2.9

Table B.3: The share of each fire station in the number of trucks used for small incidents per type-cluster.

Station \ Type-cluster	1	2	3	4	5	6	7	8	9	Avg
Aalsmeer	2.2%	0.9%	1.7%	1.8%	1.9%	1.1%	2.2%	3.0%	3.2%	2.0%
Amstelveen	7.1%	7.0%	6.3%	14.5%	5.7%	9.7%	8.7%	17.1%	8.5%	9.4%
Anton	8.2%	9.6%	6.1%	14.5%	9.5%	8.5%	7.2%	6.9%	7.8%	8.7%
Diemen	2.4%	2.2%	2.1%	2.1%	1.9%	2.2%	2.2%	2.0%	2.5%	2.2%
Dirk	5.2%	5.6%	7.5%	5.9%	7.8%	6.2%	7.6%	6.3%	8.5%	6.7%
Driemond	0.2%	0.6%	0.1%	0.0%	0.1%	0.0%	0.2%	0.2%	0.1%	0.2%
Duivendrecht	2.3%	1.8%	1.3%	6.5%	1.9%	1.2%	1.3%	0.8%	1.3%	2.1%
Hendrik	7.0%	13.5%	17.2%	8.2%	14.5%	14.0%	15.6%	7.6%	15.8%	12.6%
IJsbrand	8.5%	3.7%	3.9%	2.7%	3.8%	4.7%	4.0%	4.8%	4.3%	4.5%
Landelijk Noord	0.2%	1.6%	0.0%	0.0%	0.1%	0.1%	0.0%	0.2%	0.0%	0.3%
Nico	5.4%	9.2%	6.7%	8.6%	7.5%	8.3%	7.0%	4.8%	7.3%	7.2%
Osdorp	12.8%	8.8%	9.4%	7.5%	9.0%	9.6%	9.6%	10.7%	8.0%	9.5%
Ouderkerk a/d Amstel	0.8%	0.9%	0.3%	0.2%	0.5%	0.2%	0.6%	0.4%	0.5%	0.5%
Overig	0.3%	0.3%	0.1%	0.2%	0.2%	0.0%	0.1%	4.2%	0.1%	0.6%
Pieter	10.1%	6.6%	10.0%	8.2%	8.9%	7.2%	8.2%	10.4%	7.9%	8.6%
Teunis	6.1%	6.2%	7.4%	5.1%	6.8%	5.5%	5.1%	5.0%	5.6%	5.9%
Uithoorn	2.1%	2.0%	1.9%	2.1%	1.5%	1.7%	1.9%	2.6%	2.7%	2.1%
Victor	5.5%	6.3%	7.4%	3.5%	7.4%	6.7%	6.6%	3.0%	5.2%	5.7%
Willem	6.6%	6.3%	6.9%	4.5%	6.5%	7.6%	7.3%	4.9%	6.0%	6.3%
Zebra	7.1%	6.7%	3.6%	3.7%	4.5%	5.6%	4.6%	5.1%	4.6%	5.1%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

ESTIMATED PARAMETER VALUES IN LINEAR MODELS

Table B.4: Estimated parameter values of LM. A description of the variables can be found in Table 3.2.

		Type-cluster								
	Variable	1	2	3	4	5	6	7	8	9
β_0	Intercept	5.961	2.669	3.113	12.342	10.686	1.173	13.648	5.381	-4.851
β_1	Trend	-0.0008	0.0003	0.0006	0.0002	-0.0003	0.0008	-0.0001	-0.0007	0.0002
β_2	FG	-	0.0037	-	-	-	-	-	-	-
	FHX	-0.0115	-	-	-	-	-	-	-	-
	FXX	-	-	-	-	-	0.0011	-	-0.0081	0.0542
β_3	TG	0.0021	0.0023	-0.0003	-	-	-	0.0005	-0.0012	0.0011
	TG>0	-	-	-	-2.3572	-1.3497	0.2111	-	-	-
β_4	DR	-0.0165	-0.0070	-0.0015	-	-	-	-	-	-
	RH	-	-	-	-	-	-	-	0.0078	0.0278
	RHX	-	-	-	0.0120	-	0.0020	-	-	-
β_5	VVN	0.0019	-0.0081	-0.0043	-0.0075	-	-	-	-0.0025	-
	VVN<2	-	-	-	-	-	-	-	-	-

Table B.5: Estimated parameter values of GLM. A description of the variables can be found in Table 3.2. For the cross-terms, the same variables are used as for the single variables. For instance, for type-cluster 1, Wind*Temp. implies FHX*TG.

		Type-cluster								
	Variable	1	2	3	4	5	6	7	8	9
β_0	Intercept	6.194	2.106	3.208	11.960	10.686	4.816	13.648	5.300	-1.814
β_1	Trend	-0.0007	0.0002	0.0006	0.0002	-0.0003	0.0008	-0.0001	-0.0001	0.0002
β_2	FG	-	0.0138	-	-	-	-	-	-	-
	FHX	-0.0161	-	-	-	-	-	-	-	-
	FXX	-	-	-	-	-	-0.0313	-	-0.0097	0.0336
β_3	TG	0.0007	0.0017	-0.0020	-	-	-	0.0005	-0.0016	-0.0117
	TG>0	-	-	-	-1.9873	-1.3497	-3.3523	-	-	-
β_4	DR	-0.0093	-0.0201	0.0050	-	-	-	-	-	-
	RH	-	-	-	-	-	-	-	-0.0158	-0.1175
	RHX	-	-	-	-0.0947	-	0.9463	-	-	-
β_5	VVN	-0.0016	0.0202	-0.0082	0.0099	-	-	-	0.0170	-
	VVN<2	-	-	-	-	-	-	-	-	-
β_6	Wind*Temp.	0.0000	0.0001	-	-	-	0.0316	-	0.0000	0.0001
β_7	Wind*Rain	0.0001	0.0003	0.0000	-	-	0.0001	-	0.0001	0.0006
β_8	Wind*Visib.	0.0000	-0.0006	-	-	-	-	-	-0.0002	-
β_9	Temp.*Rain	-0.0001	-0.0001	-	0.1263	-	-0.9522	-	0.0000	0.0004
β_{10}	Temp.*Visib.	0.0001	-0.0001	0.0001	-0.0165	-	-	-	-0.0001	
β_{11}	Rain*Visib.	-0.0002	0.0003	-0.0001	-0.0005	-		-	0.0007	-

IMPORTANCE VARIABLES IN RANDOM FOREST ALGORITHM

Table B.6: The total decrease in RSS from splitting on each of these variables in the random forest algorithm.

	Total decrease in RSS per type-cluster								
Variable	1	2	3	4	5	6	7	8	9
Windspeed	3204.8	2085.4	2548.6	19659.3	13163.6	219.1	3085.7	859.9	82536.9
Temperature	9052.1	2544.0	7477.3	12254.9	19352.9	250.8	3342.0	1920.8	21008.4
Rainfall	3947.2	1332.4	3334.7	17564.5	21784.0	8.2	1259.8	1045.9	25797.8
Visibility	5908.3	1713.2	5161.2	803.9	1053.7	16.9	2072.7	1268.3	106.9

Table B.7: The frequency of being the decision variable in the root node of a decision tree in the random forest algorithm.

	Type-cluster								
Variable	1	2	3	4	5	6	7	8	9
Windspeed	0	57	302	13	86	370	327	149	496
Temperature	0	239	156	32	133	128	127	223	4
Rainfall	500	110	13	28	177	0	6	107	0
Visibility	0	94	29	427	104	2	40	21	0
Total	500	500	500	500	500	500	500	500	500

PERFORMANCE MEASURES FOR ALL MODELS

Table B.8: Quality of forecasts per type-cluster in terms of weighted mean absolute percentage error (wMAPE), including a correction by the coefficient of variation (CoV).

	wMAPE				CoV	wMAPE / CoV			
Type-cluster	LM	GLM	RF	EA*	(SD / Avg)	LM	GLM	RF	EA*
1	0.585	0.576	0.645	0.571	0.795	0.736	0.724	0.811	0.719
2	0.757	0.764	0.784	0.764	1.018	0.743	0.750	0.770	0.750
3	0.542	0.541	0.539	0.532	0.681	0.796	0.794	0.792	0.780
4	0.315	0.316	0.331	0.315	0.410	0.769	0.770	0.809	0.768
5	0.454	0.454	0.518	0.459	0.570	0.796	0.796	0.908	0.806
6	0.721	0.733	0.745	0.724	0.892	0.809	0.822	0.835	0.811
7	0.299	0.299	0.325	0.302	0.383	0.781	0.781	0.848	0.789
8	0.636	0.661	0.719	0.668	0.835	0.762	0.792	0.861	0.800
9	1.144	1.037	0.959	0.983	5.249	0.218	0.197	0.183	0.187
Avg	0.606	0.598	0.618	0.591	1.204	0.712	0.714	0.757	0.712

Table B.9: Quality of forecasts per fire station in terms of weighted mean absolute percentage error (wMAPE), including the average number of trucks used for small incidents per day.

	wMAPE				Avg #
Fire station	LM	GLM	RF	EA*	of trucks
Aalsmeer	1.051	1.052	1.051	1.051	1.4
Amstelveen	0.565	0.565	0.565	0.564	4.5
Anton	0.521	0.521	0.519	0.520	4.8
Diemen	0.951	0.951	0.946	0.949	1.3
Dirk	0.611	0.609	0.617	0.606	3.7
Driemond	1.524	1.522	1.527	1.523	0.1
Duivendrecht	2.052	2.053	2.062	2.054	0.7
Hendrik	0.458	0.453	0.462	0.450	7.3
IJsbrand	0.720	0.715	0.724	0.714	2.1
Landelijk Noord	2.029	2.035	2.044	2.037	0.1
Nico	0.563	0.560	0.559	0.557	4.3
Osdorp	0.495	0.489	0.502	0.490	5.3
Ouderkerk a/d Amstel	1.732	1.734	1.757	1.739	0.3
Overig	2.464	2.532	2.593	2.544	0.1
Pieter	0.539	0.536	0.534	0.534	5.3
Teunis	0.671	0.670	0.674	0.670	3.0
Uithoorn	1.052	1.049	1.039	1.045	1.1
Victor	0.663	0.667	0.687	0.669	2.9
Willem	0.596	0.592	0.594	0.591	3.7
Zebra	0.710	0.705	0.713	0.703	2.3
Average	0.998	1.001	1.008	1.000	2.7

RESULTS SECOND APPROACH

Table B.10: The share of each region in the number of trucks used for small incidents per type-cluster.

Region\Type-cluster	1	2	3	4	5	6	7	8	9	Avg
Buiten regio	0.8%	0.2%	0.0%	0.1%	0.1%	0.1%	0.3%	3.2%	0.2%	0.6%
Centrum	22.2%	28.9%	35.7%	21.1%	32.1%	29.8%	33.3%	26.5%	32.9%	29.2%
Haven	26.5%	15.5%	20.8%	14.7%	19.8%	20.6%	17.6%	20.0%	18.4%	19.3%
Noord	14.0%	22.1%	10.0%	11.8%	13.8%	14.1%	11.1%	12.3%	13.1%	13.6%
Oost	9.7%	8.8%	7.5%	14.9%	8.1%	7.0%	9.3%	8.4%	7.2%	9.0%
Zuidflank	10.6%	9.9%	10.3%	19.2%	7.0%	12.9%	11.9%	16.2%	12.7%	12.3%
Zuidoost	16.1%	14.4%	15.7%	18.2%	19.1%	15.4%	16.4%	13.3%	15.6%	16.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table B.11: Quality of LM forecasts per region in terms of weighted mean absolute percentage error (wMAPE), including the average number of trucks used for small incidents per day.

Region	wMAPE	Avg # trucks
Buiten regio	1.766	0.2
Centrum	0.321	16.2
Haven	0.371	10.2
Noord	0.414	7.3
Oost	0.634	4.2
Zuidflank	0.576	5.4
Zuidoost	0.370	9.0
Average	0.636	7.5

Table B.12: Quality of LM forecasts per fire station in terms of weighted mean absolute percentage error (wMAPE) for the approaches of Chapters 4 and 5 respectively, including the average number of trucks used for small incidents per day. Fire station 'Overig' is excluded here.

	wMAPE		Avg #
Fire station	LM	LM2	of trucks
Aalsmeer	1.051	0.999	1.4
Amstelveen	0.565	0.560	4.5
Anton	0.521	0.522	4.8
Diemen	0.951	0.956	1.3
Dirk	0.611	0.608	3.7
Driemond	1.524	1.549	0.1
Duivendrecht	2.052	1.854	0.7
Hendrik	0.458	0.457	7.3
IJsbrand	0.720	0.726	2.1
Landelijk Noord	2.029	2.248	0.1
Nico	0.563	0.562	4.3
Osdorp	0.495	0.496	5.3
Ouderkerk a/d Amstel	1.732	1.670	0.3
Pieter	0.539	0.541	5.3
Teunis	0.671	0.665	3.0
Uithoorn	1.052	1.024	1.0
Victor	0.663	0.651	2.9
Willem	0.596	0.600	3.7
Zebra	0.710	0.706	2.3
Average	0.921	0.915	2.9

ALLOCATION TABLES

Table B.13: Allocation of prediction per region over all fire stations according to the allocation rule in Equation 10.

Fire station\Region	Buiten regio	Centrum	Haven	Noord	Oost	Zuidflank	Zuidoost
Aalsmeer	5.9%	2.1%	3.0%	1.7%	1.6%	5.2%	1.7%
Amstelveen	5.9%	4.7%	4.5%	3.6%	5.0%	11.3%	4.1%
Anton	4.5%	2.7%	3.1%	4.0%	4.7%	4.0%	9.9%
Diemen	4.4%	3.0%	3.4%	5.4%	5.4%	3.8%	10.5%
Dirk	5.8%	11.6%	7.0%	6.8%	7.2%	6.0%	4.6%
Driemond	3.6%	1.9%	2.3%	2.8%	2.5%	2.8%	4.6%
Duivendrecht	4.9%	3.8%	3.9%	5.7%	9.2%	4.7%	10.4%
Hendrik	5.8%	12.3%	8.5%	6.2%	5.0%	5.4%	3.8%
IJsbrand	5.1%	5.6%	7.3%	6.5%	3.7%	3.9%	3.5%
Landelijk Noord	3.7%	2.3%	2.9%	4.7%	2.9%	2.7%	4.4%
Nico	5.2%	6.6%	5.9%	10.8%	7.8%	4.7%	5.7%
Osdorp	7.1%	4.7%	9.0%	2.7%	2.4%	5.1%	2.2%
Ouderkerk a/d Amstel	5.5%	4.0%	3.9%	3.7%	5.8%	7.8%	5.1%
Pieter	6.8%	9.4%	8.4%	3.7%	3.7%	7.2%	3.0%
Teunis	5.9%	9.3%	10.0%	5.1%	3.8%	4.9%	3.3%
Uithoorn	5.6%	2.3%	3.1%	1.9%	1.9%	7.4%	2.1%
Victor	4.9%	4.8%	4.7%	9.1%	9.1%	4.5%	8.3%
Willem	5.3%	5.9%	5.1%	7.7%	14.3%	5.4%	7.2%
Zebra	4.2%	3.1%	3.8%	7.8%	4.0%	3.2%	5.5%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Table B.14: Allocation of prediction per region over all fire stations according to the empirical shares of the fire stations in each region.

Fire station\Region	Buiten regio	Centrum	Haven	Noord	Oost	Zuidflank	Zuidoost
Aalsmeer	17.5%	0.0%	0.0%	0.0%	0.0%	9.1%	0.0%
Amstelveen	2.1%	0.9%	0.0%	0.0%	0.1%	75.4%	0.0%
Anton	19.2%	0.0%	0.0%	4.8%	0.6%	0.1%	50.3%
Diemen	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	12.4%
Dirk	0.0%	23.6%	0.0%	0.3%	2.2%	0.1%	0.0%
Driemond	0.0%	0.0%	0.0%	0.0%	1.0%	0.0%	0.1%
Duivendrecht	0.0%	0.0%	0.0%	0.0%	29.4%	0.0%	1.0%
Hendrik	0.0%	39.9%	0.8%	7.0%	0.0%	0.0%	0.0%
IJsbrand	4.2%	0.0%	20.4%	1.3%	0.0%	0.0%	0.0%
Landelijk Noord	3.1%	0.0%	0.0%	1.7%	0.0%	0.0%	0.0%
Nico	0.0%	0.5%	2.0%	48.5%	0.6%	0.0%	2.8%
Osdorp	19.2%	0.7%	46.7%	0.0%	0.0%	0.0%	0.0%
Ouderkerk a/d Amstel	0.0%	0.0%	0.0%	0.0%	5.2%	0.0%	0.0%
Pieter	33.2%	30.4%	1.0%	0.0%	0.2%	0.3%	0.0%
Teunis	1.0%	1.3%	29.0%	0.0%	0.0%	0.0%	0.0%
Uithoorn	0.3%	0.0%	0.0%	0.1%	0.0%	14.9%	0.0%
Victor	0.0%	0.0%	0.0%	0.8%	1.0%	0.0%	32.0%
Willem	0.0%	2.7%	0.0%	0.0%	59.8%	0.1%	1.3%
Zebra	0.0%	0.0%	0.0%	35.3%	0.0%	0.0%	0.0%
Total	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

BIBLIOGRAPHY

[1] Brandweer Amsterdam-Amstelland. Organisatiemodel. [Accessed December 7, 2016]. (Cited on page 1.)

[2] Previous research done first by Daphne van Leeuwen and Rob van der Mei (CWI) and later by Ab Boersema (as Research Paper BA). Both are not published, but known at fire brigade Amsterdam-Amstelland. (Cited on page 2.)

[3] Gemeente Amsterdam - Onderzoek, Innovatie & Statistiek. Basisbestand Gebieden Amsterdam (BBGA), October 2016. [Accessed November 7, 2016].

[4] G Lenderink, GJ van Oldenborgh, E van Meijgaard, and J Attema. Intensiteit van extreme neerslag in een veranderend klimaat. *Meteorologica*, 20(2):17–20, 2011. (Cited on page 12.)

[5] Julian J Faraway. Practical regression and anova using r., 2002. p.39–41. (Cited on page 35.)

[6] Brandweer Amsterdam-Amstelland. Kazernes en locaties. [Accessed November 8, 2016].