Severe Weather-based Fire Department Incident Forecasting

Guido Legemaate*, Jeffrey de Deijn†, Sandjai Bhulai‡ and Rob van der Mei‡§

*Fire Department Amsterdam-Amstelland, Ringdijk 98, 1097 AH Amsterdam
Email: g.legemaate@brandweeraa.nl
†Vrije Universiteit Amsterdam, Faculty of Science, De Boelelaan 1081, 1081 HV Amsterdam
Email: jeffreydeijn@hotmail.com
‡Vrije Universiteit Amsterdam, Department of Mathematics, De Boelelaan 1081, 1081 HV Amsterdam
Email: s.bhulai@vu.nl
§Centrum Wiskunde en Informatica, Science Park 123, 1098 XG Amsterdam
Email: mei@cwi.nl

***Abstract*—For fire departments, having enough firefighters available during a shift is obviously an important requirement. Nevertheless, just like in any organization, having too many firefighters standby is not desirable from a financial point of view. Despite the fact that fire departments can and should not be run like production companies, at least for staffing purposes, forecasting the number of incidents that each fire station has to handle is highly relevant. In this paper, we develop models to create a forecast for the number of incidents that each fire station in the Dutch safety region Amsterdam-Amstelland has to handle for specific incident types and deal with major and small incidents. Previous studies mainly focused on multiplicative models containing correction factors for the weekday and time of year. Our main contribution is to incorporate the influence of different weather conditions in the categories of wind, temperature, rain, and visibility. Rain and wind typically have a strong linear influence, while temperature mainly has a non-linear influence. We show that an ensemble model has the best predictive performance.**

Keywords–incident forecasting; fire department planning; generalized linear models; ensemble models; severe weather conditions.

I. INTRODUCTION

As for most organizations, the ability to accurately forecast demand is of "paramount importance" for emergency services, fire departments included [1][2]. In the 1970s, the Fire Department of the City of New York and The New York City-Research And Development (RAND) Institute jointly conducted various groundbreaking studies [3]. More recent academic interest seems to be focused more on ambulance services. While there are obvious similarities between emergency service providers, they differ in (the number of) incident types, demand characteristics, and operational logistics.

Nevertheless, the problems that fire departments have to deal with, like loss of coverage and the degradation of response times, are similar. The same is true for possible gains. At a strategic and tactical level, improved forecasting of workload leads to a better placement of base stations, and improved staffing and scheduling. At an operational level, one may pro-actively relocate units to maximize coverage and minimize response times during major incidents [4]. All things considered, efficient planning of emergency service resources is crucial.

Demand is an important factor when models are being developed to improve the performance of emergency service providers. It is, however, not uncommon that, for instance, call arrival rates are estimated using adhoc or rudimentary methods such as averages based on historical data [5]. This may ultimately lead to a degradation of performance, or over- or under-staffing [6]. In most cases, reducing response times is an important performance measure since this increases the survival rate of victims [7][8].

Numerous papers have been written on forecasting forest or wildfire occurrences, many of those using weather variables and vegetation types as part of their model [9]. Forest fire forecasting is no longer a study in academia alone. In fact, in the United States, e.g., the National Interagency Coordination Center operates a predictive service which provides decision support to the United States Forest Service, which facilitates proactive management and planning of fire assets on both operational and tactical levels [10].

Although the scale of wildfire occurrences in the Netherlands is smaller than in many other parts of the world, it is mainly the greater interrelationship of different types of infrastructure, i.e., the wildland-urban interface, that causes concern and even lead to surface fuel models for the Netherlands [11]. For a more urban environment, like the conurbation of Western Holland, which also includes Amsterdam, forest fire occurrences are not very common.

The occurrence of certain types of incidents which fire departments in urban settings typically respond to also correlate with weather conditions. As such, incorporating this information into the planning process of emergency services yields important advantages over current practice. Typical weather and storm-related incidents that fire departments in the Netherlands respond to are fallen trees, potentially falling debris that needs securing (roofs, construction work, scaffolding), and water damage. Another important factor is that the weather also impacts fire department operations by overwhelming available resources.

At least in the Netherlands, to the best of our knowledge, there are no known applications of forecasting algorithms that are used in practice at fire departments, being urban or specialized forest services. Given this, we aim to provide an easily applicable model that can be put to use for a general fire department when dealing with severe weather conditions. Therefore, we quantify and model the fact that - under these conditions - fire departments experience an increased amount of incidents, which in itself leads to an increased amount of deployments.

The organization of this paper is as follows. In Section II, we describe the data used to obtain the forecasts. Section III describes the models used for forecasting. In Section IV, we analyze the performance of the models and state the insights. Finally, in Section V, we conclude and address a number of topics for further research.

II. DATA

The available data contains one row for each incident that happened in the region Amsterdam-Amstelland from January 2008 up until April 2016. The most interesting information includes the incident's start- and end time, location, incident type, the concerned fire station, and the number of fire trucks used. Since the size of incidents matters for the number of people you need, the focus is on forecasting the number of trucks needed.

A. *Major and small incidents*

The vast majority of incidents require only one or otherwise just a few trucks. Therefore, it makes sense to distinguish between 'major' and 'small' incidents. Major incidents are mostly due to coincidences that are hard to predict. Specifically, they do not rely on bad weather conditions or a particular time of the year in the Netherlands, for example, as with forest fires in countries with a tropical climate. This arouses the expectation that the inter-incident times of major incidents can be modeled as a Poisson process.

To test the Poisson assumption, we apply the Kolmogorov-Smirnov (KS) test on the inter-incident times in cases when more than k trucks are needed for several values of k. The KS-test shows that if we define an incident as 'major' when at least $k = 6$ trucks are used, then the KS-test does not reject exponentially of the inter-incident times (approximate p-value $= 0.429$). However, for values of $k < 6$, the KS-test doubts (or rejects) this exponentially (approximate p-value $= 0.073$ and 0.002 when at least $k = 5$ and $k = 4$ trucks are used, respectively. Hence, according to this result, we define an incident to be major when at least six trucks are needed.

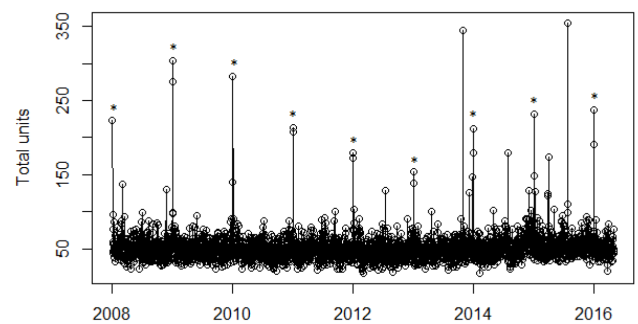


Figure 1. Total number of trucks used for small incidents per day. * *Peaks caused due to an increased amount of incidents around New Year's Eve.*

Next, we focus on the small incidents. Small incidents are probably easier to predict, since bad weather conditions often cause many *small* incidents to happen (like fallen trees, water damage, or police/ambulance assistance at traffic accidents). To study this, we first omit all incidents on December 31 and January 1. There are extremely many incidents around New Year's Eve as can be seen on Figure 1, mainly caused by fireworks-related incidents. These conditions do not occur in the rest of the year, therefore we model these days separately as described in the modeling section.

After elimination we find that not all outliers in Figure 1 are New Year's days. In fact, the only five days that, for the amount of trucks used per day (>138), on par with New Year's day are days with severe weather conditions as can be seen in Table I.

On these days with severe weather conditions only 0.46%, instead of an average 1.72%, of incidents are major incidents. Without a clear reason to assume that the frequency of major incidents on this particular type of days is lower, there must be another explanation rather than chance. If so certain circumstances cause many small incidents to happen, like those caused by severe

TABLE I. WEATHER CONDITIONS ON THE FIVE DAYS THAT COULD COMPETE WITH NEW YEAR'S DAYS IN TERMS OF AMOUNT OF TRUCKS USED.

Date (day)	Trucks	Highest windspeed (km/h) Overall	Highest windspeed (km/h) Worst hour	Total rainfall (mm) Overall	Total rainfall (mm) Worst hour
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

weather conditions. Data from the fire department on incident types that happened on days with severe weather conditions further support this finding.

B. Seasonal patterns

There are clear seasonal patterns in the data for the number of trucks needed throughout each year, week, and day. The plots in Figure 2 illustrate this. The pattern in Figure 2c depicts the activity cycle that an average person goes through every day of the week. The week pattern (Figure 2b) differs per type of incident and looks a little different throughout the year. The pattern in Figure 2a 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. We correct for this by using a Loess-smoothed function over the factors per week. We will include all these patterns in our model.

C. Weather variables

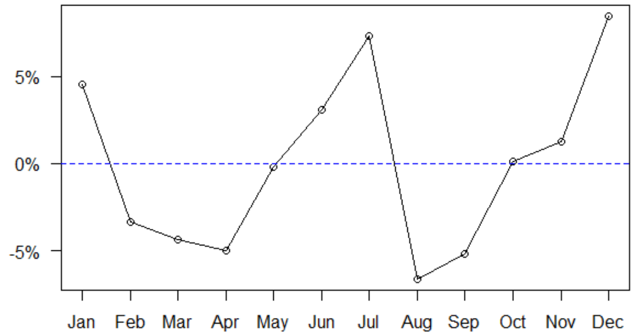
Besides the time-dependent components, we want to know which weather variables we must include in our model. Therefore, we use the Pearson correlation test to determine which weather conditions have a significant influence on the number of trucks we need. The results of these tests are summarized in Table II.

TABLE II. PEARSON'S PRODUCT-MOMENT CORRELATION TESTS BETWEEN SOME WEATHER VARIABLES AND THE NUMBER OF TRUCKS USED FOR SMALL INCIDENTS PER DAY.

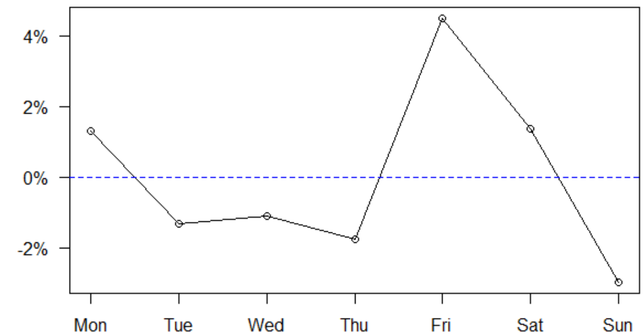
Category	Variable	p-value	Correlation
Wind	Average wind speed (FG)	$< 10^{-12}$	0.132
	Maximum hourly mean wind speed (FHX)	$< 10^{-15}$	0.177
	Maximum wind gust (FXX)	$< 10^{-15}$	0.189
Temperature	Average temperature (TG)	0.6897	0.007
	Boolean: 1 if average > 0 (TG>0)	$< 10^{-8}$	0.105
Rainfall *	Rainfall 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.*

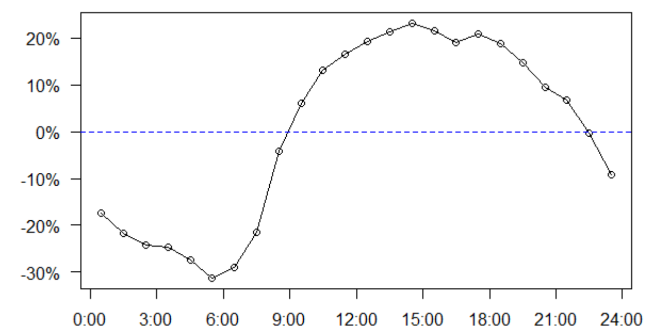
We can see from this that the minimum visibility and the average temperature both have no significant (direct)



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



(b) Week pattern: peak on Friday.



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

Figure 2. Seasonal patterns: the given percentages represent relative differences with respect to the average (in blue).

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.

D. Fireworks-related incidents

It is a tradition in the Netherlands to celebrate New Year's Eve with fireworks. Only then, the general public is allowed to light fireworks. Fireworks need to comply with legal standards, and may only be sold during the last three days of the year at licensed shops.

Over the years, fire departments in the Netherlands have seen a slow but steady rise in fireworks-related incidents [12]. Most common incident types that fire departments respond to during New Year's Eve are dumpster fires, outside fires and vehicle fires. Also more serious incidents happen, like in 2020 a fire in an Arnhem flat that left two members of a family dead, and two other family members critically injured. The fire began in the ground floor hallway of a high-rise apartment building and was identified to be caused by fireworks. A family of four were found trapped in an elevator, which shut down as the building lost electricity due to the fire. Noteworthy, but not related to New Year's Eve, is the Enschede fireworks disaster of May 13, 2000. A catastrophic explosion in a fireworks depot, situated in a residential area of the eastern Dutch city of Enschede, essentially obliterated the neighborhood of Roombeek [13].

In 2014, in an attempt to mitigate the nuisance caused by fireworks-related incidents on New Year's Eve, the Dutch government reduced the time window in between the public was allowed to set off fireworks. This reduced time window was set to 6 pm on December 31 to 2 am on January 1, while before it was allowed starting from 10 am on December 31.

Figures 3 and 4 show the average number of trucks used for small incidents per hour around New Year's Eve before and after New Year's Eve 2014/2015, respectively. The reduced time window, and possible relation between fireworks and small incidents, seems to be reflected in the average number of trucks used per hour as well. Furthermore, it seems that the reduction has only compressed all incidents into a smaller time window, as the average total number of trucks per hour has increased at certain periods. These preliminary findings however need further research to find out whether this is a coincidence or not.

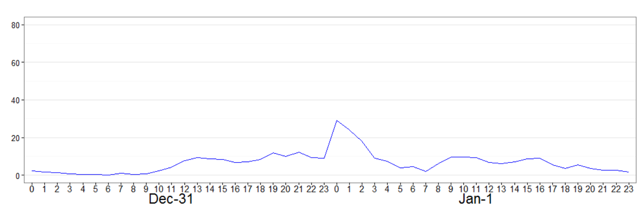


Figure 3. Average number of trucks used for small incidents per hour around New Year's Eve before 2014/2015.

III. Models

In this section, we will create a model that predicts directly the number of trucks that each fire station needs.

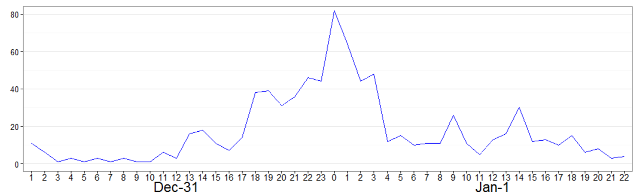


Figure 4. Average number of trucks used for small incidents per hour around New Year's Eve after 2014/2015.

In the previous section, we have shown that the major incidents (with at least six trucks needed) are very hard to predict and that we can best model them by an (inhomogeneous) Poisson process. We also 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.

Fire departments in general have a variety of incident types they respond to. Not all of them occur frequently enough to make a good forecast on. Since these in this aspect have little value, they are eliminated and the remaining incident types are clustered based on their correlation with certain weather variables.

TABLE III. INCIDENT CLUSTERS AND CORRELATION WITH RESPECT TO WIND SPEED, TEMPERATURE, RAINFALL, AND VISIBILITY.

Cluster	Type	Wind	Temp.	Rain	Visib.	# p/day
1	Outside fire	-0.135	0.09	-0.193	0.075	3.46
2	Animal in water	-0.088	0.134	-0.058	0.013	1.65
	Animal assistance	-0.072	0.129	-0.088	0.069	
	Person in water	-0.041	0.056	-0.023	0.009	
	Locked out	-0.006	0.159	-0.043	0.062	
3	Contamination / nuisance	-	-0.228	0.038	-0.111	2.52
4	Locked in elevator	-	-0.088	0.021	-0.015	8.16
	Automated alarm	-	-0.069	0.051	-0.037	
5	Fire rumor	-	-0.103	-	-	3.57
	Inside fire	-	-0.038	-	-	
	General assistance water	-	-0.019	-	-	
6	Police assistance	0.048	-0.062	0.026	-	1.34
7	Ambulance assistance	-	-0.065	-	-0.039	8.55
	Vehicle in water	-	-0.042	-	-0.025	
	Reanimation	-	-0.086	-	-0.008	
8	General assistance	0.063	0.079	0.057	0.052	2.28
9	Storm- and water damages	0.319	0.028	0.279	-	2.10

In total, we now have nine different incident clusters in our dataset, some of which occur much more/less often than others. In Table III, we show the correlation with respect to one variable of each four weather categories. Looking at these correlations in detail, we can see that these are often in line with our expectations. For instance, high wind speed and rainfall obviously increase the number of incidents due to 'storm and water damage' (type 9) and decrease the likelihood of 'outside fires' occurring (type 1).

We will estimate, for each incident type t, a model that predicts the number of trucks used for *small* incidents $y_{t,d}$ on date d, i.e.,

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

Here, $f_{t,d}$ is a correction factor for the week number based on a Loess-smoothed function as in Figure 5, and $g_{t,d}$ is a weekday factor as in Figure 2b. Both are computed separately for each incident type. 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.

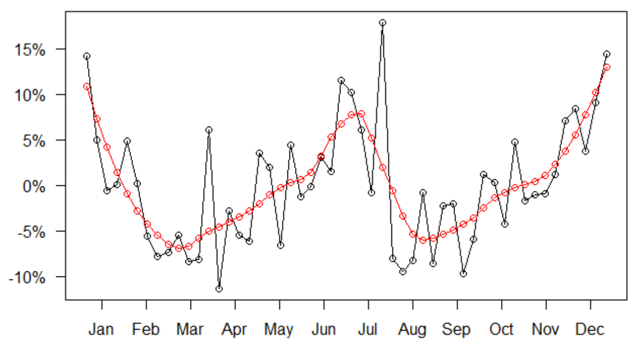


Figure 5. The year pattern per week (in black) together with its Loess-smoothed variant ($\alpha = 0.3$).

A. Linear regression model

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

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

where $\epsilon_{t,d}$ is assumed to have expectation zero and some finite variance. Note that this model includes an intercept (β_0), a linear trend ($\beta_1 \cdot d$) and (at most) four weather variables.

B. Generalized Linear Model

Our second model, a Generalized Linear Model (GLM) arises from an observation that the largest outlier neither has the highest wind speed nor the most rainfall. However, the *combination* of wind and rainfall might be the cause. It may, therefore, be a good idea to include also cross-effects in our model, i.e.,

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

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.

C. Random Forests

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.

As input, the algorithm needs a $T \times (K + 1)$-matrix with K explanatory variables 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 is drawn with replacement from the input matrix. On this sample, a decision tree (DT) algorithm is executed. This procedure is repeated N times, yielding N decision trees. When a new sample comes in, we can take all N predictions for this sample and average these to get the final prediction.

D. Performance measures

To evaluate the different models, we create a train and a test set. The train set contains all data up until 2015/06. The test set contains all data from 2015/07 onwards. This holds for all incident types, 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 on n samples using the Mean Absolute Percentage Error (MAPE), assuming $y_t > 0$,

$$\mathrm{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \widehat{y}_t}{y_t}\right| = \frac{1}{n}\sum_{t=1}^{n}\frac{|y_t - \widehat{y}_t|}{y_t},$$

as well as its weighted version, i.e.,

$$\mathrm{wMAPE} = \frac{\sum_{t=1}^{n}\frac{|y_t - \widehat{y}_t|}{y_t} y_t}{\sum_{t=1}^{n} y_t} = \frac{\sum_{t=1}^{n}|y_t - \widehat{y}_t|}{\sum_{t=1}^{n} y_t}.$$

Here, y_t is the true value in time period t and $\widehat{y}_t$ is the prediction.

E. Fireworks-related modeling

All models from the previous sections are based on data without both major incidents and all incidents on New Year's Eve. Since New Year's Eve, in terms of amount of small incidents, is far from normal, we can not just make a forecast for those days with the current models. Using an ensemble method, a forecast for all occurrences of New Year's Eve in our dataset was made. With the real amount of trucks used subtracted from this, we assume that this result approximates the number of fireworks-related incidents.

Table IV shows the correlation between some weather variables and fireworks-related incidents. These incidents occur more often on a cold New Year's Eve with little wind and rain.

TABLE IV. CORRELATION BETWEEN WEATHER VARIABLES AND FIREWORKS-RELATED 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

Besides the fact that our dataset only holds 17 New Year's Eves, we also found out that due to policy changes the time window for setting off fireworks has been changed. Therefore, only two New Year's Eves in our dataset are completely representative for future ones. Based on these limitations we implement a simple linear model with just an intercept and four weather variables.

The results of estimating the model on all New Year's Eves are given in Table V, including p-values of two-sided t-tests to test the null hypothesis that the true parameter equals zero.

If we estimate the model just on the first twelve New Year's Eves and leave the last five for testing, we get a MAPE of 0.349. Due to too little New Year's Eves in our dataset we are unable to make a very accurate forecast in this particular occasion. This may very well also be the reason for the lack of significant predictive power by the weather variables. Since New Year's Eve from many different perspectives is not a regular day, certainly agreed upon by the fire department, we chose to use this simple estimation thus not to spend more time trying to improve upon this model.

TABLE V. PARAMETER ESTIMATES OF A LINEAR MODEL FOR FIREWORKS-RELATED INCIDENTS.

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

IV. RESULTS

In this section, we will compare the performance of the different models and evaluate the insights derived from them. The results on the MAPE and wMAPE values are given in Table VI. These performance measures are based on the total daily number of trucks used for small incidents (over all fire stations and types). This enables us to compare all models through one value. It is also interesting to see how significant a parameter is on a 1 to 5 scale, as in Table VII for LM, Table VIII for GLM, and Table IX for RF. Here, we assign 1 when the p-value < 0.001 (very significant) until 5 when the p-value ≥ 0.1 (not significant).

TABLE VI. PERFORMANCE MEASURES OF THE MODELS.

Model	MAPE	wMAPE
LM	0.1886	0.1924
GLM	0.1865	0.1880
RF	0.2006	0.2019

A. Linear regression model

For the linear model, comparing Table VII to Table III, we observe that when a weather variable has significant predictive power for some type, then their mutual correlation is relatively high as well. This is a nice result, but unfortunately, the reverse is not true. For instance, type 3 is highly correlated with one of the temperature variables, but this variable does not have predictive power for this type, which is surprising.

If we look at Table VII in more detail, it stands out that several types have no weather variables with significant

TABLE VII. SIGNIFICANCE OF ESTIMATED PARAMETERS FOR LM.

Variable	Type cluster (see Table III)) 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
Wind speed	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:* $p < 1$

predictive power. Opposed to type 3, this is not surprising for types 6 and 7, since their correlations to the weather variables are relatively low as well. On the other hand, types 1 and 9 are well predicted by the amount of wind and rainfall, which is intuitively explainable as well.

Since the wMAPE is higher, we 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.

B. *Generalized Linear Model*

Recall that the GLM model is an expanded version of the linear model, so it could be at least as good. The question is how much value it adds to the linear model. Comparing the significance of the variables in Table VIII to that of LM in Table VII, we observe that, in general, the single weather variables have lost some importance in favor of cross-term variables they partition in. Type 1 is an excellent example of this. Here, the 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 wind speed and rainfall turn out to be less predictive on their own than the LM indicated. It is their cross-term effect that is important. Looking at the average column on the right, we also see that the intercept has lost some importance. Apparently, a bigger part can be modeled by the weather after adding some cross-term variables. Of all weather variables, it is even the case that two cross-term variables have the most predictive power.

Noting the influence of the cross-term variables, we expect that the performance of the GLM is better than that of the LM. If we compute the results for the totals per day, we still see that the wMAPE is somewhat higher than the MAPE, but compared to their equivalents of the LM, they are slightly better (about 2%).

TABLE VIII. SIGNIFICANCE OF ESTIMATED PARAMETERS FOR GLM.

Variable	Type cluster (see Table III)) 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
Wind speed	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:* $p < 1$

TABLE IX. IMPORTANCE W.R.T. TOTAL DECREASE IN RSS.

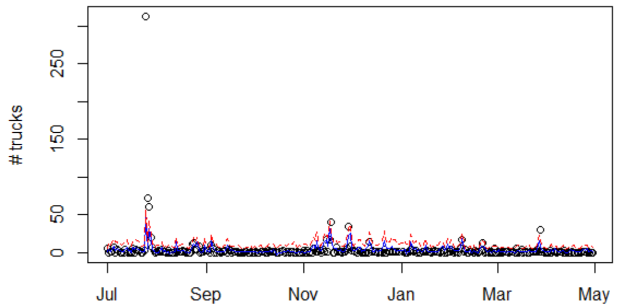
Variable	Type cluster (see Table III)) 1	2	3	4	5	6	7	8	9	Avg
Wind speed	4	2	4	1	3	2	2	4	1	2.56
Temperature	1	1	1	3	2	1	1	1	3	1.56
Rainfall	3	4	3	2	1	4	4	3	2	2.89
Visibility	2	3	2	4	4	3	3	2	4	3.00

C. *Random Forests*

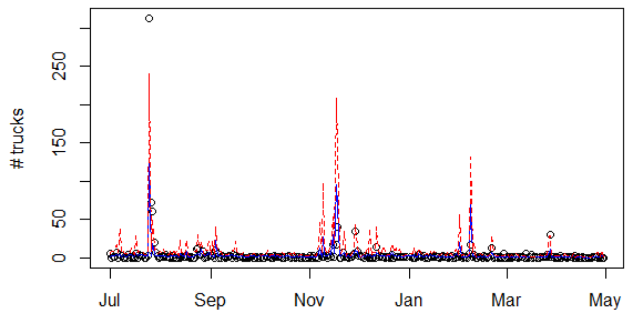
Different from the previous models, the RF algorithm does not estimate a parameter for each variable. We, therefore, have to find another measure for the importance of each variable. We will consider the 'RSS-ranking' for this purpose.

In the RF algorithm, 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. Table IX shows the results of the RSS ranking. 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.

When we compare the results of RF to the previous models, we see that, in general, RF gives the worst results. However, the effort for running this model is perhaps not in vain. When diving deeper into the results, we discover that the RF has the best wMAPE for type 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 9 of both GLM and RF in Figure 6. Obviously, the RF algorithm recognizes much better than GLM when the weather conditions are risky and likely to cause many incidents to happen.



(a) Generalized Linear Model



(b) Random Forest

Figure 6. Forecasts (in blue) of the number of trucks used for small incidents of type 9, including the upper bound of its 95%-prediction interval (in red).

D. Ensemble model

From the previous discussion, we 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. Motivated by this, we propose to use *ensemble averaging* (EA), defined by

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

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

We have to determine the optimal 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 $\gamma < 0.5$. When we vary γ from 0 to 1, both the MAPE = 0.1853 and the wMAPE = 0.1860 take their minimum in $\gamma^* = 0.2$ (which is better than GLM individually; when compared with $\gamma = 0$).

TABLE X. CAPACITY NEEDED PER DAY AND FIRE STATION WITH CERTAINTY THIS CAPACITY SUFFICES THAT DAY.

	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

E. Practical implication

After the forecasts are complete, we extract from them 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.

The $100(1 - \alpha)$%-prediction interval for the GLM model $y = X^\intercal \beta + \epsilon$ for a future observation y_0 can be computed as [14]

$$\widehat{y}_0 \pm t_{n-k}^{(1-\alpha/2)} \widehat{\sigma} \sqrt{x_0^\intercal (X^\intercal X)^{-1} x_0 + 1},$$

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, and k is the number of variables in the model.

For the RF algorithm, we have 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.

If we combine all these results, we get Table X that gives the needed capacity for each fire station. 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. Then 'Amstelveen' also needs a capacity of two on some days. Moreover, 'Pieter' does not have the required capacity in 1.7% of the days (see in red).

V. Conclusions and discussion

In this paper, we developed a model to create a 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.

The answer is split into two parts. The forecasts created for the small incidents can be done reasonably well by EA. Major incidents can be modeled by an inhomogeneous Poisson process. Concerning the weather, (the combination of) rain and wind on average had the 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 has the least predictive power among those four weather variables.

The current implementation computes a different model per type-cluster and subsequently divides the total prediction over the fire stations. One enhancement for further research that certainly seems logical is to make an estimate per region instead of per station. Incidents happen at a certain place in a certain region, not where a truck of a certain fire station happened to be in the vicinity of. An added benefit is that we can more accurately use the characteristics of separate regions. For example, a region with many big and/or old trees may be more at risk during days with severe weather conditions. To expand on that even further, this risk can subsequently be adjusted for seasons (e.g., spring vs. winter) and/or regions (e.g., different tree types) with more or less leaves on the trees, making them respectively more or less prone to falling over due to wind gusts.

Some of these characteristics can be captured by first dividing the prediction per type-cluster over all regions according to a certain weight. As a proof of concept we calculated the share of each region in the number of trucks used for small incidents per type-cluster, of which the results can be found in Table XII. Using the LM of Section III-A we calculated the results per region, as shown in Table XI. As expected, we find that - in terms of wMAPE - when fewer incidents happen, it gets harder to make a good forecast. When we calculate the totals per day we observe that MAPE(LM2) = 0.1887 and wMAPE(LM2) = 0.1919, which is in line with our previous finding for the results of the LM as shown in Table VI. Note that in both cases we used the same models for the type-clusters, which not surprisingly led to similar results. Future research should be able to generate a model which can be applied to each separate region, while still taking all different incident types into account, and come up with a way to divide the prediction over all fire stations.

TABLE XI. QUALITY OF LM FORECASTS PER REGION IN TERMS OF WMAPE AND AVERAGE NUMBER OF TRUCKS USED FOR SMALL INCIDENTS PER DAY.

Region	wMAPE	Avg # trucks
External	1.766	0.2
Center	0.321	16.2
Harbor area	0.371	10.2
North	0.414	7.3
East	0.634	4.2
South	0.576	5.4
Southeast	0.370	9.0
Average	0.636	7.5

References

[1] G. A. G. Legemaate, S. Bhulai, and R. D. van der Mei, "Applied urban fire department incident forecasting," in Proceedings of IARIA Data Analytics, Sep. 2019.

[2] J. B. Goldberg, "Operations Research Models for the Deployment of Emergency Services Vehicles; EMS Management Journal," EMS Management Journal, vol. 1, no. 1, 2004, pp. 20–39.

[3] J. M. Chaiken and J. E. Rolph, "Predicting the demand for fire service," RAND Corporation, P-4625, 1971.

[4] P. L. van den Berg, G. A. G. Legemaate, and R. D. van der Mei, "Increasing the responsiveness of firefighter services by relocating base stations in Amsterdam," Interfaces, vol. 47, no. 4, 2017, pp. 352–361.

[5] D. S. Matteson, M. W. McLean, D. B. Woodard, and S. G. Henderson, "Forecasting emergency medical service call arrival rates," The Annals of Applied Statistics, vol. 5, no. 2B, 2011, pp. 1379–1406.

[6] H. Setzler, C. Saydam, and S. Park, "EMS call volume predictions: A comparative study," Computers & Operations Research, vol. 36, no. 6, Jun. 2009, pp. 1843–1851.

[7] M. P. Larsen, M. S. Eisenberg, R. O. Cummins, and A. P. Hallstrom, "Predicting survival from out-of-hospital cardiac arrest: a graphic model." Annals of emergency medicine, vol. 22, no. 11, Nov. 1993, pp. 1652–8.

[8] M. Gendreau, G. Laporte, and F. Semet, "The Maximal Expected Coverage Relocation Problem for Emergency Vehicles," The Journal of the Operational Research Society, vol. 57, 2006, pp. 22–28.

[9] A. Ganteaume, A. Camia, M. Jappiot, J. San-Miguel-Ayanz, M. Long-Fournel, and C. Lampin, "A Review of the Main Driving Factors of Forest Fire Ignition Over Europe," Environmental Management, vol. 51, no. 3, Mar. 2013, pp. 651–662.

TABLE XII. 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
External	0.8%	0.2%	0.0%	0.1%	0.1%	0.1%	0.3%	3.2%	0.2%	0.6%
Center	22.2%	28.9%	35.7%	21.1%	32.1%	29.8%	33.3%	26.5%	32.9%	29.2%
Harbor area	26.5%	15.5%	20.8%	14.7%	19.8%	20.6%	17.6%	20.0%	18.4%	19.3%
North	14.0%	22.1%	10.0%	11.8%	13.8%	14.1%	11.1%	12.3%	13.1%	13.6%
East	9.7%	8.8%	7.5%	14.9%	8.1%	7.0%	9.3%	8.4%	7.2%	9.0%
South	10.6%	9.9%	10.3%	19.2%	7.0%	12.9%	11.9%	16.2%	12.7%	12.3%
Southeast	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%

[10] N. I. C. C. U.S.A. Predictive Services Program Overview. Last accessed on 11/30/2021. [Online]. Available: https://www.predictiveservices.nifc.gov

[11] B. P. Oswald, N. Brouwer, and E. Willemsen, “Initial Development of Surface Fuel Models for The Netherlands,” Forest Research: Open Access, vol. 06, no. 02, 2017.

[12] Brandweer Nederland. Jaarwisseling: incidenten voor het tweede jaar op een rij toegenomen. Last accessed on 11/30/2021. [Online]. Available: https://www.binnenlandsbestuur.nl/bestuur-en-organisatie/nieuws/brandweer-moest-vaker-uitrukken.11915233.lynkx

[13] P. G. van der Velden, C. J. Yzermans, and L. Grievink. New York, NY, US: Cambridge University Press, 2012, ch. Enschede Fireworks disaster, pp. 473–496.

[14] J. J. Faraway, “Practical regression and ANOVA using R,” University of Bath, 2002.