
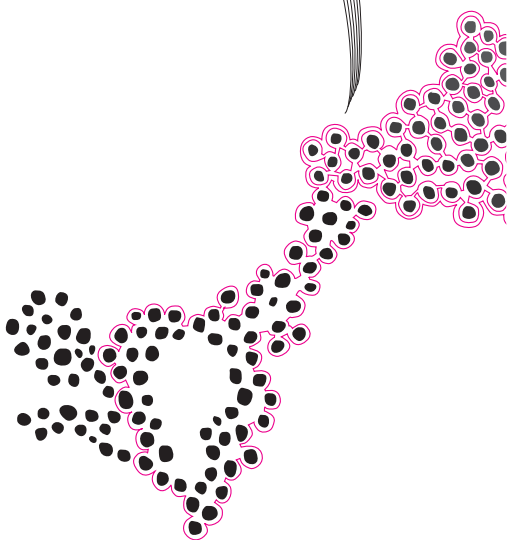


BSc Thesis Applied Mathematics



Data-driven kitchen fire prediction based on environmental variables

Dorien van Leeuwen



Supervisor: Maurits de Graaf, Marie-Colette van Lieshout, Changqing Lu

July, 2022

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

Data-driven kitchen fire prediction based on environmental variables

Dorien van Leeuwen

July, 2022

Abstract

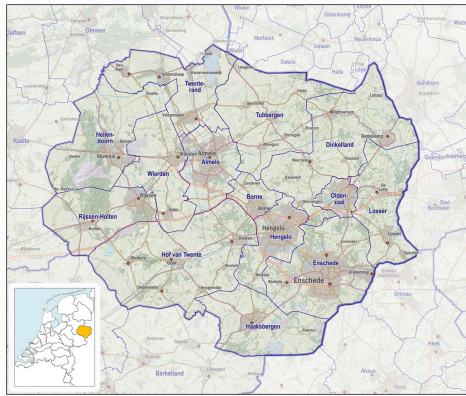
Efficient prediction is crucial to preventing harm caused by kitchen fires. In this paper, we propose a kitchen fire model using the data collected by the Twente Fire Brigade. Specifically, we utilize the permutation techniques of random forests and perform classic stepwise regression methods to select the explainable environmental variables. For unstable results, we propose stabilization methods. Moreover, we build a Poisson generalized linear model which successfully captures the spatial patterns seen in the data.

Keywords: fire prediction, poisson GLM, random forest, stepwise regression, variable importance

1 Introduction

Building fires often have a big impact on the people involved. Accurate prediction can help firefighters to prevent harm. The Twente Fire Brigade handles more than 2.000 fire incidents annually [2]. To improve their service, the Twente Fire Brigade has been interested in data-driven fire risk management research and an collaboration with the University of Twente was started.

The Twente Fire Brigade has provided the data of fire incidents, which has been used previously in research on chimney fire incidents [5, 6, 7]. We will follow a similar approach to that of chimney fire prediction from Lu et al [5] but now for the prediction of kitchen fires. Our goal is to provide an accurate model using the least amount of variables.



(A) The map of Twente



(B) Location of kitchen fire incidents

FIGURE 1: map of Twente with all towns and cities (a) and spatial projection of the reported kitchen fire incidents in the years 2004 - 2020 (b)

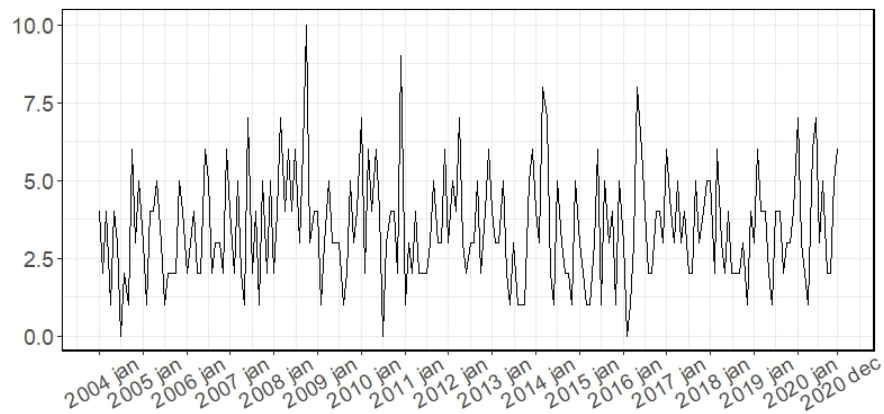


FIGURE 2: Temporal projection of the reported kitchen fire incidents during 2004-2020

Totally, there are 694 kitchen fire incidents from the years 2004-2020. Of these fire incidents, we have access to the time (figure 2) and the location (figure 1) on several levels of accuracy, as precise as latitude and longitude and in categories such as area boxes and municipalities. In figure 1, a concentration in the cities is visible, especially the bigger ones: Almelo, Hengelo and Enschede. In figure 2, we see no apparent time dependency or seasonal effect.

In this paper, we build a predictive model based on a selection of environmental variables. In Section 2, we introduce the data collection on environmental variables. In Section 3, we employ random forests [1] and stepwise regression [3] methods to select explanatory variables. In Section 4, we develop a Poisson General Linear Model and validate the model using residuals. In section 5, we conclude and talk about future work.

2 Data

Aside from the fire incident data, we also collect areal unit data for 500 by 500 meter boxes in the Twente region, see table 1. From IFV ¹, we access the building information, aggregated building year and function type. From CBS ², we access information about the population and density. Additionally, after discussion with firefighters, compared to chimney fires, we obtain extra variables from CBS, variables 23 - 27 from table 1, that appear to be closely related to kitchen fires.

TABLE 1: Environmental variables with their abbreviation, description and source

Variable	Abbreviation	Description	Source
V_1	House	The total number of houses	IFV
V_2	House_indu	The number of houses with an industrial function	IFV
V_3	House_hotl	The number of houses with an hotel function	IFV
V_4	House_resi	The number of houses with an residential function	IFV
V_5	House_20	The number of houses constructed before 1920	IFV
V_6	House_2045	The number of houses constructed between 1920 and 1945	IFV
V_7	House_4570	The number of houses constructed between 1945 and 1970	IFV
V_8	House_7080	The number of houses constructed between 1970 and 1980	IFV
V_9	House_8090	The number of houses constructed between 1980 and 1990	IFV
V_{10}	House_90	The number of houses constructed after 1990	IFV
V_{11}	House_frsl	The number of free standing houses	IFV
V_{12}	Resid	The number of residents	CBS
V_{13}	Resid_14	The number of residents with an age in the range of 0 till 14	CBS
V_{14}	Resid_1524	The number of residents with an age in the range of 15 till 24	CBS
V_{15}	Resid_2544	The number of residents with an age in the range of 25 till 44	CBS
V_{16}	Resid_4564	The number of residents with an age in the range of 45 till 64	CBS
V_{17}	Resid_65	The number of residents with an age of 65 or higher	CBS
V_{18}	Man	The number of male residents	CBS
V_{19}	Woman	The number of female residents	CBS
V_{20}	Address	The density of addresses in the box	CBS
V_{21}	Urbanity	The urbanity of the block	CBS
V_{22}	Town	Boolean variable indicating the presence of a town	CBS
V_{23}	Poor	The percentage of poor residents (income 0-20 percent)	CBS
V_{24}	Rich	The percentage of rich residents (income 80-100 percent)	CBS
V_{25}	Value_house	The average value of the houses in the block	CBS
V_{26}	Gas_use	The average gas use in m^3 in the block	CBS
V_{27}	Elec_use	The average electricity use in kWh in the block	CBS

Income and the value of the house can influence the state of the kitchen, since people with disposable income could spend it to improve their kitchen and expensive houses have better quality kitchens. Income is divided into 5 categories, each containing 20 percent of the Dutch population. People categorized as '**Poor**' have an income in the bottom 20 percent of the income range. People categorized as '**Rich**' are in the top 20 percent income class. Poor and Rich are percentages of the population in a box, that belong to either income class.

¹IFV: Instituut Fysieke Veiligheid

²CBS: Centraal Bureau voor de Statistiek

The variables `'Gas_use'` and `'Elec_use'` are the averages for gas and electricity use in the box. The choice between gas and electricity might also be important for fire risk, as there could be a difference in fire occurrences, but there is also a correlation between utility use and time at home.

Overall, we consider these newly added variables because they might be correlated with latent variables `'state of kitchen'` and `'amount of kitchen use'`.

3 Variable Selection

3.1 Methods

Of our 27 available environmental variables, we make a selection to include in the model. Not all variables will be significantly correlated with kitchen fires and some variables are mutually dependent. Hence it will be undesirable for the model performance to include all environmental variables. We compare two different methods to select explanatory variables, a currently used method and a more classic method. First, we use permutation techniques from random forests, as this was successful for the chimney fire prediction [5]. Second, we use stepwise regression, as it is a more classic statistical method which we want to compare to random forests.

3.1.1 Random Forests

The permutation importance is defined to be the decrease in a model score when a single variable is randomly permuted. The variables with the most decrease are deemed important with this method. Strobl et al [9] suggest using conditional permutation importance to prevent bias towards correlated variables in the importance scores.

We implement the random forest and permutation importance techniques with the `'party'` package from R as suggested by Strobl et al [8]. The trees are conditional inference trees [4], which are unbiased in the candidate selection for each node, giving all variables an equal appearance in the random forest. The chimney fire prediction [5] using this approach was successful.

The function `'cforest'` from the `'party'` package was used with unbiased control and 1000 trees. For different `mtry` (figure 7 in Appendix), the number of variables considered at each node, we perform conditional permutation to obtain the variable importance.

3.1.2 Stepwise regression

Next, we look at stepwise regression for a comparison. With stepwise regression, on each step a variable is either added or removed depending on the method and the criteria. The criteria we use is the Akaike Information Criteria (AIC). In forward selection, at each step a variable is added if the criteria is satisfied, thus if the AIC decreases. In backwards elimination, the process starts with a full model and removes variables until the criteria does not improve anymore. These can also be combined into stepwise selection, where forward selection and backwards elimination are performed simultaneously.

Generally, backwards elimination is preferred, but with more complex models with a lot of variables, forward selection is still possible [3]. The function `step()` from package `'stats'` with direction `'both'`, a combination of `'forward'` and `'backward'`, a form of stepwise selection, was used to determine the best model by the Akaike Information Criteria. This selects 12 variables which were then ordered based on the p-value in table 2.

3.1.3 Preliminary results

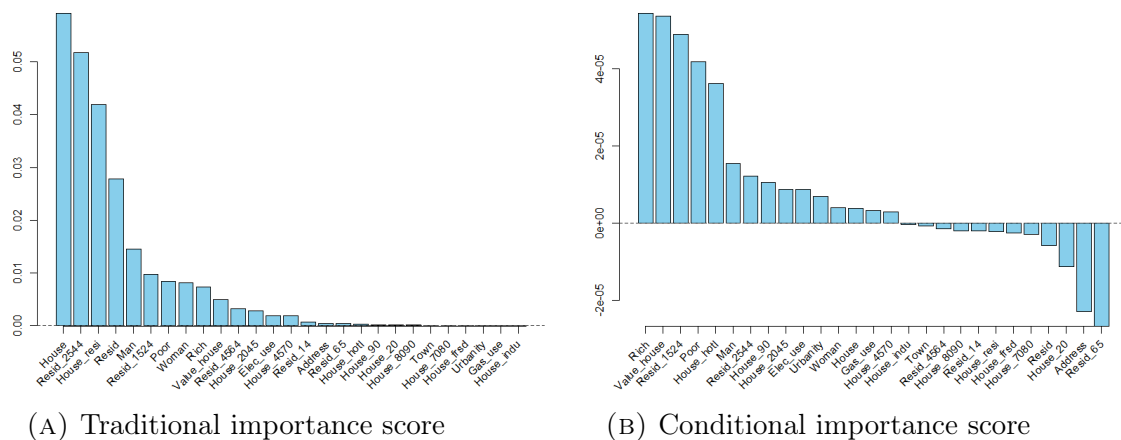


FIGURE 3: Variable importance scores from Random Forest

The preliminary results of the random forest, in figure 7 in the Appendix, contain a lot of variation between the different mtry. According to Strobl et al [8], the variability of the conditional importance is lower than that of the traditional importance within each level of mtry.

However, we still unfortunately need to conclude that for now the results are unstable. In the results with different mtry, shown in the Appendix, figure 7, variables do not have a consistent position; for example, the variable `'Value_house'` has the first position for mtry 24, second position for mtry 9, but last and very negative position for mtry 18.

TABLE 2: Variable importance from stepwise regression

Variable	p value
Poor	< 2e-16
Town	3.76e-14
House	1.42e-05
House_resi	3.60e-05
Resid_65	4.22e-04
Resid	0.0018
Gas_use	0.005
House_90	0.036
Resid_4564	0.058
House_indu	0.071
House_frsd	0.072
Resid_1524	0.133

The results of the stepwise regression, table 2, would more easily compare to the traditional importance score, figure 3a, which we are not using because of potential bias towards correlated variables. Also, the variable '*Town*', scores very high on the stepwise regression, but is not deemed important by both the traditional and conditional importance scores.

Overall, an comparison is difficult and with the known issues of stepwise regression, such as bias towards correlated variables and being less effective with more potential explanatory variables, we continue with stabilizing the random forest results.

3.2 Stabilization

To stabilize the importance scores, we perform two methods; iterative random forest and averaging the importance scores.

3.2.1 Iterative random forest

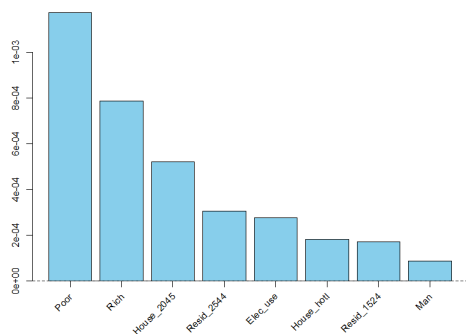
In this process, seen in figure 8 in the appendix, we determine the positive importance scores based on the random forest results and repeat the process with only the positive variables. The process is completed when all remaining variables have an positive importance score.

3.2.2 Averaging of importance scores

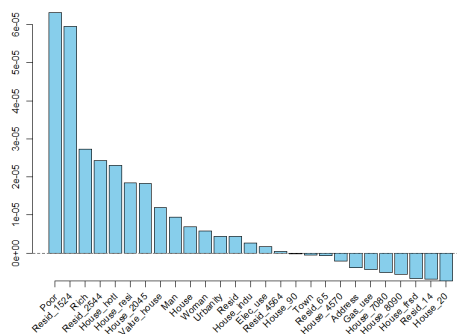
The random forest tests with different mtry (figure 7) included at each node show unsteady results. Averaging these results lead to a more stable conclusion; variables that have an high importance score at any particular or multiple

metry, will also have an higher averaged importance score. This lessens the variation of the random forests, which makes the important variables stand out more.

3.2.3 Results



(A) End result for iterative random forest



(B) Average importance scores

FIGURE 4: Variable importance scores from Random Forest

3.3 Important variables

Looking at both the iterative random forest and the averaging of importance scores, figures 4a and 4b, these variables are selected as tentatively important, with hypothetical explanations for their inclusion:

- V_{23}, V_{24} - Poor and Rich: Income could be correlated to the state of the kitchen, and inclusion of both poor and rich can show an wider range in income division.
- V_{14}, V_{15} - Young/Adult: The people who are most likely to cook, and young people might be less experienced and more easily distracted while cooking.
- V_6 - Old houses: The age of the house could be related to the age of the kitchen, and the piping and wiring in the house could be inadequate for modern requirements.
- V_{27} - Electricity use: Many fires start with an electrical appliance, and it is possibly also correlated with how often people are at home and potentially cooking.
- V_4 - Residential houses: Buildings that contain an kitchen and where people usually cook.

- V_{18} - Men: Probably an correlation to the amount of people in general, as the variable woman was eliminated late in the iterative random forest. However, it could also be possible that men are just more likely to have an kitchen fire.

The variable hotel also shows up, but there are very little buildings with an hotel function, and thus also almost no fire incidents in hotels, see table 3. Even if there would be an difference in fire risk, the influence on the final fire prediction would be insignificant.

TABLE 3: Categories of houses of fire incidents in the data

	Hotel	Other	Total
Old	0	15	15
New	4	675	679
Total	4	690	694

In total, we have 8 variables to consider for the model. If we separately consider the inclusion of all variables, we need to consider too many different models. Thus, the variables we always include in the model are the number of poor residents, the number of rich residents, the young people (age 15-24) and the old houses (build between 1920 and 1945), because these are the main important variables as indicated by the stabilization. The variables we consider to further include are the adults (age 25-44), electricity use, number of residential houses and men.

4 Model

4.1 Model formulation

We assume that the N_i are independent and Poisson distributed, and that the expected number of kitchen fires in a box is proportional to the number of houses in the box. The model will look like this:

$$N_i \sim \text{Poisson}(h_i \lambda) \quad (1)$$

where N is the number of fire incidents in box i , h_i the number of houses in box i , and the intensity function is given by:

$$\lambda = \exp(\theta_1 + \theta_2(V_{23} * V_{12}) + \theta_3(V_{24} * V_{12}) + \theta_4 V_6 + \theta_5 V_{14} + \theta_6 V_{27} + \theta_7 V_4 + \theta_8 V_{18}) \quad (2)$$

which includes the variables V_{23} : poor residents, V_{24} : rich residents, V_6 : old houses, V_{14} : young residents, V_{27} : electricity use, V_4 : residential houses and V_{18} : men.

TABLE 4: Basic model ($V_{23}, V_{24}, V_6, V_{14}$) and additional variables sorted by AIC

Included variables	AIC
BASIC + $V_{27} + V_4 + V_{18}$	2367.373
BASIC + $V_{15} + V_{27} + V_4 + V_{18}$	2367.540
BASIC + $V_{27} + V_{18}$	2367.737
BASIC + $V_{15} + V_{27} + V_{18}$	2368.232
BASIC + $V_{15} + V_{27} + V_4$	2402.890
BASIC + $V_{15} + V_{27}$	2407.062
BASIC + $V_{27} + V_4$	2414.807
BASIC + V_{27}	2424.157
BASIC + $V_{15} + V_{18}$	2446.227
BASIC + $V_{15} + V_4 + V_{18}$	2446.749
BASIC + V_{18}	2448.886
BASIC + $V_4 + V_{18}$	2449.784
BASIC + $V_{15} + V_4$	2504.733
BASIC + V_{15}	2508.080
BASIC + V_4	2518.402
BASIC	2526.310

These variables are included in the model, because they were indicated by the importance scores of the random forest results. Sorting the models we consider by their AIC score, table 4, we see that next to the basic variables, we also include V_{27} : electricity use, V_4 : residential houses and V_{18} : men.

4.2 Model fitting

After fitting the model using the data from 2004-2017,

TABLE 5: Parameters fitting of model

Parameter	Estimate
θ_1	-4.032
θ_2	4.629e-04
θ_3	-3.092e-04
θ_4	-9.569e-06
θ_5	-4.188e-03
θ_6	3.019e-04
θ_7	7.138e-04
θ_8	4.851e-03

we plot the predicted spatial projections versus the actual kitchen fires in the testing years to visualize the results. In 2018 and 2019 there were 36 and 40 fires respectively.

The predictions are almost the same since a year is a short time for any spatial data to change and we did not have access to the change in the number of houses, however, there is a slight difference in total predicted fires, 38.21 in 2018 and 38.17 in 2019.

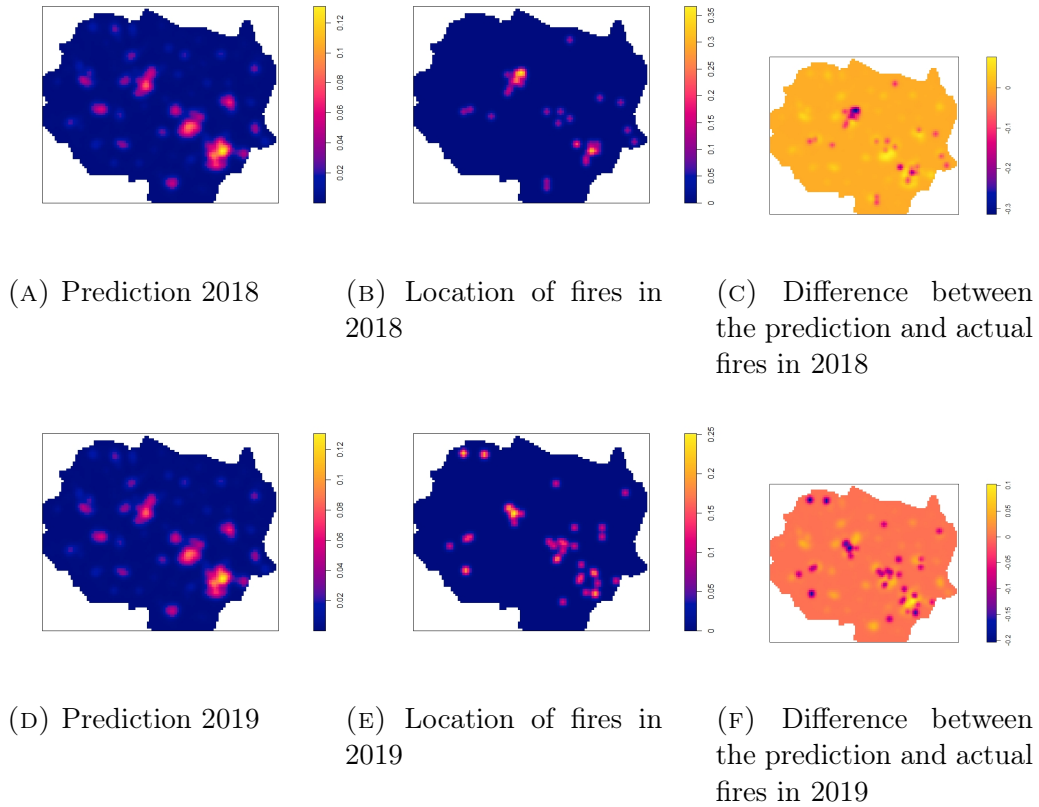


FIGURE 5: Model prediction compared to the actual fire incidents

The predictions compared to the testing years look reasonable, especially for the year 2019. The sum of the absolute difference is 70.25888 for 2018, and 73.46695 for 2019. In the difference squared, which penalizes larger errors, the prediction for 2019 is better with 39.52603 versus 39.79313 for 2018. Due to noise, however, the model will never be an perfect match for the actual fire instances, and especially with the low amount of fires, on average 40 a year, randomness always has an visible influence.

4.3 Model validation

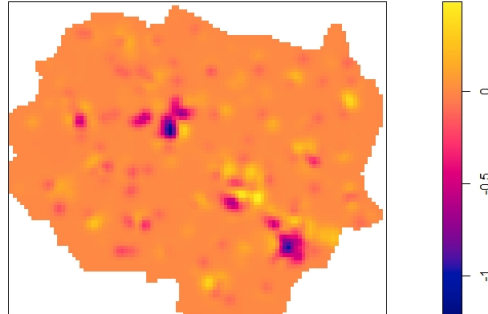


FIGURE 6: Residuals

When we look at the residuals, we see that our model underestimates for parts of the bigger cities, Almelo, Hengelo and Enschede. The model fits for the entire region of Twente and it is possible the characteristics of the cities are not adequately captured by the model. For instance, cities have an on average poorer population. The sum of the residuals, however is with -0.03322588 , really small.

5 Conclusion

We needed 7 variables to obtain a good performing model, which is more than initially expected. However, since the occurrence of kitchen fires is dependent on human behaviour and we can not access the latent variable 'state of kitchen', the need for this amount of variables seems logical.

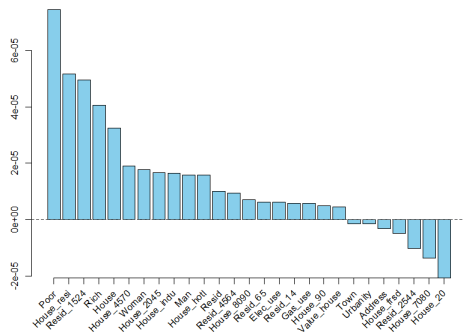
As advise to the firefighters, we have seen in all methods and results that the variable 'poor' is crucial in the kitchen fire prediction. Prevention strategy should be focused on the poorer neighbourhoods. Old houses seem slightly less likely to have an kitchen fire, but this effect is small when we look at the parameters of the fitted model.

For future research, an expansion towards the more continuous Poisson point process model should be considered, to obtain an more accurate and precise model. There is also potential for more research into stabilization of the random forest results and the reasons for the instability; for instance, which stabilization method gives better results and methods on how to quantify (un)stability.

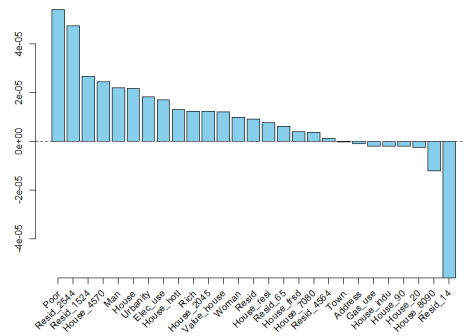
References

- [1] L. Breiman. Random forests. In *Machine Learning*, pages 5–32, 2001.
- [2] CBS. Branden en hulpverleningen; alarmering van de brandweer, regio 2013-2019, 2022.
- [3] G. Heinze, C. Wallisch, and D. Dunkler. Variable selection - a review and recommendations for the practicing statistician. *Biometrical Journal*, 60, 01 2018.
- [4] T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, 15(3):651–674, 2006.
- [5] C. Lu, M.-C. van Lieshout, M. de Graaf, and P. Visscher. Chimney fire prediction based on environmental variables. 2021. 63rd World Statistics Congress, ISI 2021 ; Conference date: 11-07-2021 through 16-07-2021.
- [6] C. Lu, M.-C. van Lieshout, M. de Graaf, and P. Visscher. Data-driven chimney fire risk prediction using machine learning and point process tools. Workingpaper, arXiv.org, 2021.
- [7] M. School, M. de Graaf, M.-C. van Lieshout, E. Sanders, and R. de Wit. Van tellen naar voorspellen: Sturen op risico's met een voorspellend wiskundig model op basis van historische brandweerdeata. *Tijdschrift voor veiligheid*, 20(1):60–74, Jan. 2021.
- [8] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1), 2008.
- [9] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1), 2007.

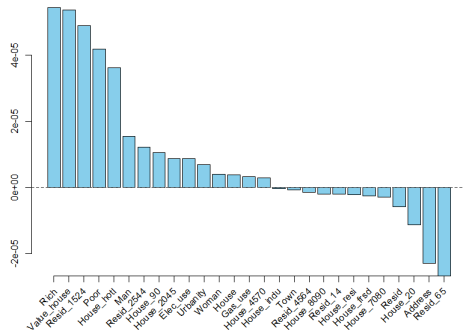
Appendix



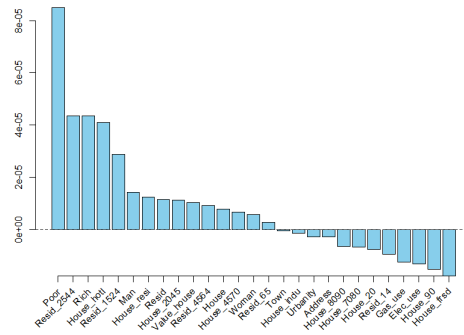
(A) $mtry = 3$



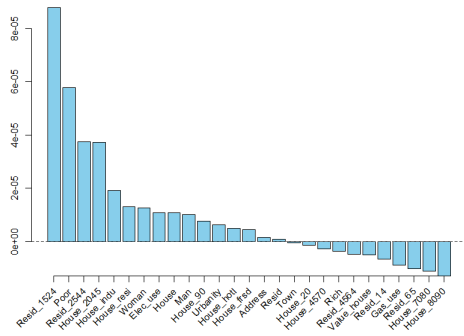
(B) $mtry = 6$



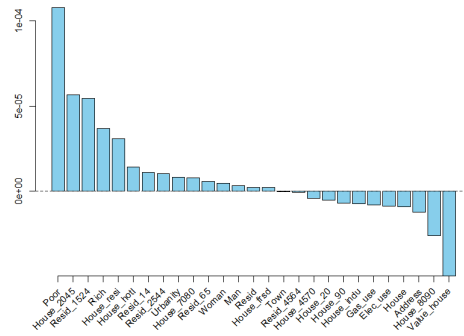
(C) $mtry = 9$



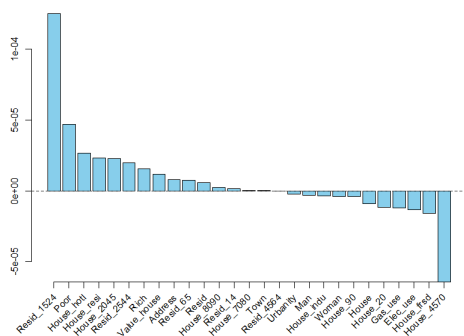
(D) $mtry = 12$



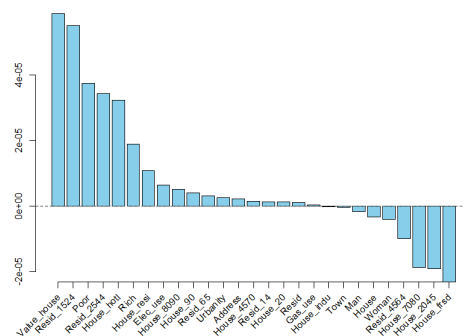
(E) $mtry = 15$



(F) $mtry = 18$



(G) $mtry = 21$



(H) $mtry = 24$

FIGURE 7: Conditional importance results from RF with different values for $mtry$, the number of variables considered at each node

