Chimney fire prediction based on environmental variables

Changqing Lu*[1], Marie-Colette van Lieshout[2,1], Maurits de Graaf[3,1], Paul Visscher[4]

[1]University of Twente, Enschede, The Netherlands

[2]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

[3]Thales Nederland B.V., Huizen, The Netherlands

[4]Brandweer Twente, Enschede, The Netherlands

Abstract

Fire prediction and prevention is closely related to public security. In this paper, we propose a chimney fire prediction model using data collected by the Twente Fire Brigade. Random forests are used to select explanatory variables from a large set of candidates. Furthermore, we build a nested Poisson generalized linear model based on house types and weather variables. Plausible predictions are obtained, that capture the salient spatial and temporal patterns seen in the data.

Keywords— areal unit data, conditional variable importance, Poisson generalized linear model

1 Introduction

The Dutch fire and rescue services are developing an increasing interest in business intelligence as part of their strategy for data driven fire risk management (NVBR, 2010). In order to be able to take appropriate counter measures against fire, precise risk prediction should be performed. As an initial step, the Twente Fire Brigade has enhanced their data collection, resulting in structured databases containing detailed information on the incidents of various fire types from 2004 to the present. The ultimate goal is to construct an accurate risk prediction system for fire services via appropriate mathematical models. Here, we concentrate on one of the most commonly occurring fire types – the chimney fire.

Fire risk is usually influenced by environmental variables, such as geographical information (e.g., geography and urbanity) and temporal information (e.g., weather). In this work, we combine techniques from machine learning and statistics. Random forests (Breiman, 2001) are employed to select the most important environmental variables. Based on that, we build a Poisson generalized linear model that can be used for formal inference and prediction. The model is validated on data from the Twente region. Benefiting from the data-driven selection of explanatory variables, we can utilize the information contained in environmental variables so as to design an appropriate model accordingly.

2 Related Work

There is early literature on fire prediction based on point process models. Møller & Diaz-Avalos (2010) employed a shot-noise Cox process and Serra et al. (2014) considered a log-Gaussian Cox process to model the risk of wild fires. Peng et al. (2005) used the risk index proposed by the Los Angeles fire brigade as explanatory variable to define fire risk. Specific for chimney fires, as a pilot study for our research, School (2018) also employed a log-Gaussian Cox process to predict the risk considering both environmental and latent factors. These references mainly explored appropriate models for fire prediction but did not focus on the selection of explanatory variables for their prediction model. To perform a comprehensive data-driven study for fire prediction, we employ machine learning techniques to select the most relevant environmental variables in a non-parametric way before designing our prediction model.

3 Data Collection

Collaborating with the Twente Fire Brigade, we collected the data of 2050 reported chimney fire incidents[1] occurring between Jan 1, 2004, and Dec 31, 2019, in the Twente region (map shown in Fig. 1a). Each incident is reported individually with its ID, location, time and a brief description on the basis of which an incident is classified as a chimney fire or not. The spatial and temporal projections are shown in Fig. 1b and Fig. 1c. It is clearly visible in the figures that the spatial distribution is heterogeneous in the sense that most reported chimney fires occur in urban areas, and the temporal distribution is periodic since chimney fires occur more in the winter than in the summer.

E-mail addresses of the authors: c.lu@utwente.nl (Changqing Lu), marie-colette.van.lieshout@cwi.nl (Marie-Colette van Lieshout), m.degraaf@utwente.nl (Maurits de Graaf) and p.visscher@brandweertwente.nl (Paul Visscher).

[1]An automatic retrieval of chimney fires was performed here using queries for specific words in the unstructured text descriptions of the alarm centre. Manual verification may lead to a different result, because sometimes misreported chimney fires can occur.

In addition, 29 environmental variables (spatial: 23, temporal: 6) that may influence chimney fire occurrences were collected as well. The spatial variables cover the building, population and urbanity information and are collected over $500m \times 500m$ pre-defined square boxes in Twente. The temporal variables record the daily weather data collected at the weather station at Twente airport, as well as seasonal information. Specific abbreviations, descriptions and sources of the variables are listed in Table 1. To assess the influence of small variations in weather among different parts of the Twente region, we analyzed the data from two neighbouring weather stations. Similar relations with the chimney fire occurrences were obtained, which illustrates that those small variations can be ignored. To unify the counting of the incident data and all environmental data, we also counted the incidents over $500m \times 500m$ pre-defined square boxes and days. Doing so, we obtain structured areal unit data consisting of 6291 boxes $\times$ 5840 days, where all 29 environmental variables are accessible for every box and every day.

(a) Map of Twente (b) Spatial pattern (c) Monthly counts

Figure 1: Map of the Twente region (a) and spatial (b) and temporal (c) projections of the reported chimney fire incidents during 2004–2019.

4 Selection of Explanatory Variables

Obviously, not all 29 environmental variables are significantly correlated with chimney fires and some of them are mutually dependent. Hence, it is unnecessary and undesirable to include them all in a prediction model. To select the important ones, we perform variable importance analyses on spatial and temporal variables respectively using a machine learning technique, random forest (Breiman, 2001). As non-parametric models, random forests fit the candidate variables to the response variable and measure the importance of a variable via randomly permuting its values across all tree-predictors and reporting the increase of the prediction error before and after the permutation. In our implementation, we use unbiased random forests (Strobl et al., 2007) to avoid the random forest construction biasing towards factorial variables with many categories or continuous variables with many cut points. Additionally, considering the high correlation between certain variables (e.g., urbanity and population), we use the conditional permutation importance (Strobl et al., 2008) instead of the traditional permutation importance to suppress the variable importance biasing towards correlated variables. Our implementation of the variable importance analysis is based on Debeer & Strobl (2020), which provides a faster computation and shows more stable results than the original method (Strobl et al., 2008).

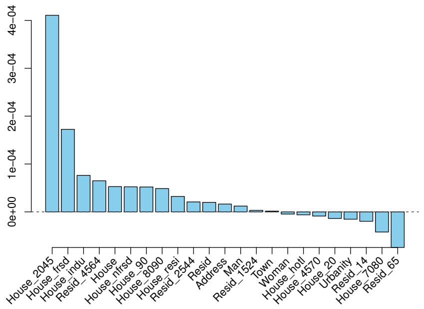
The conditional permutation importance results obtained for spatial and temporal variables are plotted in Fig. 2. The vertical axis refers to the increase of the model prediction error when the conditional permutation on a variable is applied. A large increase indicates that the variable under permutation is very important for correct predictions, while a decrease indicates that the variable has no influence on or even hampers the prediction. With this technique, the number of houses constructed between 1920 and 1945 (House_2045) and the number of freestanding houses[2] (House_frsd) are shown to be the most important spatial variables, and the wind speed (WindSpeed) and wind chill (WindChill) are shown to be the most important temporal variables. The results could have the following intuitive explanations. Chimney pipes in old buildings tend to be made of brick rather than metal, which increases the risk to catch a chimney fire. Secondly, most freestanding houses contain chimneys whereas other types do not. For temporal variables, strong wind can fuel a fire, and wind chill reflects people feeling cold thus inducing them to use their fires and chimneys.

5 Nested Fire Prediction Model

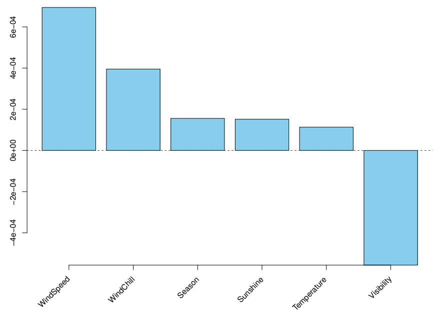
With the selected explanatory variables, we proceed to build a fire prediction model. Since we found that the type of a house could affect its risk to catch a chimney fire, we divide the houses into four groups depending on the age (constructed between 1920 and 1945 or not) and on whether or not they are freestanding. We assume that, within a group, the expected number of chimney fires is proportional to the number of houses of that type. The risk of a house to catch a chimney fire is type-dependent and subject to seasonal variations as well as to wind speed and wind chill. Finally, we assume that the random variables $N_{i,t}^k$, the number of chimney fires in box i on day t in houses of type k, are independent and Poisson distributed. These considerations suggest the following model:

$$N_{i,t} = \sum_k N_{i,t}^k \sim Poisson(\sum_k h_i^k \lambda_t^k) \qquad k = 1, ..., 4, \tag{1}$$

[2]Freestanding houses here refer to detached and semi-detached houses.



(a) Spatial variable importance



(b) Temporal variable importance

Figure 2: Conditional permutation importance of spatial and temporal variables.

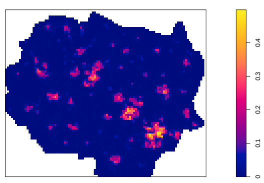
where $N_{i,t}$ indicates the number of chimney fires occurring in box i on day t, h_i^k indicates the number of house of type k in box i and λ_t^k indicates the fire risk intensity for a house of type k on day t. The intensity function is designed as

$$\lambda_t^k = exp(Harmonic^k(t, o_h^k) + Polynom^k(WindChill, o_{p1}^k) + Polynom^k(WindSpeed, o_{p2}^k)), \tag{2}$$

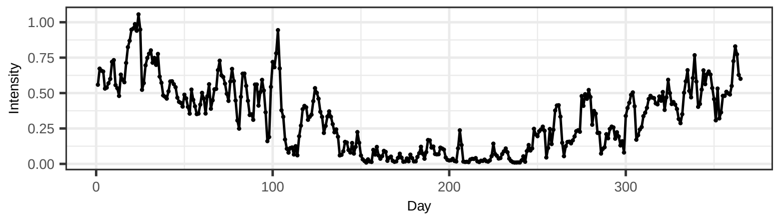
where a harmonic function[3] with order o_h^k is employed to model seasonal variations, two polynomial functions with order o_{p1}^k and o_{p2}^k are used to fit wind chill and wind speed and the exponential function guarantees that the risk intensity function stays positive. The parameters of our model are estimated via maximum likelihood estimation using data over the whole Twente region during 2004–2018. To determine the appropriate function orders (i.e., o_h^k, o_{p1}^k o_{p2}^k) simultaneously, we select the combination having the smallest Akaike information criterion. This way, it is found that, from temporal variables, only wind chill needs to be included in the prediction model. Hence, the final intensity function becomes

$$\lambda_t^k = exp(Harmonic^k(t, o_h^k) + Polynom^k(WindChill, o_p^k)). \tag{3}$$

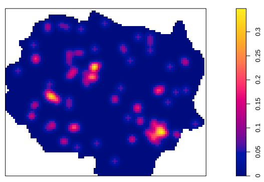
To validate our model, we test it on the data of 2019. The spatial and temporal predictions as well as the actual realisations are plotted in Fig. 3. Overall, our model obtains good performance on both spatial and temporal predictions. Spatially, it concentrates on the urban areas containing considerable numbers of houses. Temporally, it complies with the detailed trends observed in actual realisations.



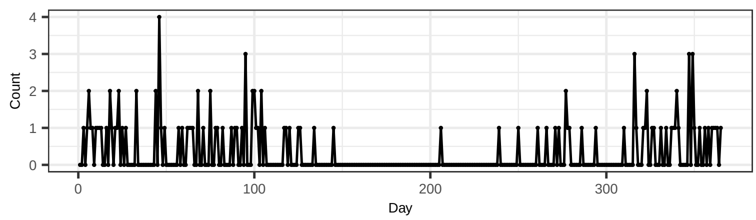
(a) Spatial predictions



(b) Temporal predictions



(c) Spatially smoothed actual incidents



(d) Temporal actual counts

Figure 3: Predictions and actual realisations for the year 2019 visualized spatially and temporally.

[3]Harmonic function here refers to consine and sine functions.

6 Conclusion

In this paper, we developed a model to predict chimney fire incidents based on environmental variables and showed that plausible predictions could be obtained capturing the features of the data. In future, we aim to refine our model in two directions. Firstly, we will fit the observed spatio-temporal point pattern (cf. Fig. 1b) directly. Secondly, we will embed the model in a hierarchical framework that allows for interaction and random effects.

Acknowledgements

This work is funded by the Dutch Research Council (NWO) for the project "Data Driven Risk Management for Fire Services" (18004). We thank the user committee for their valuable input.

References

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Debeer, D., & Strobl, C. (2020). Conditional permutation importance revisited. *BMC Bioinformatics*, *21*, 307.

Møller, J., & Diaz-Avalos, C. (2010). Structured spatio-temporal shot-noise Cox point process models, with a view to modelling forest fires. *Scandinavian Journal of Statistics*, *37*(1), 2-25.

NVBR. (2010). De brandweer over Morgen. *Arnhem: Nederlandse Vereniging voor Brandweer en Rampenbestrijding*.

Peng, R., Schoenberg, F., & Woods, J. (2005). Time conditional intensity model for evaluating a wildfire hazard index. *Journal of the American Statistical Association*, *100*(469), 26-35.

School, M. (2018). A log-Gaussian Cox process for predicting chinmey fires at fire department Twente. *MSc thesis, University of Twente*.

Serra, L., Saez, M., Mateu, J., Varga, D., Juan, P., Diaz-Avalos, C., & Rue, H. (2014). Spatio-temporal log-Gaussian Cox process for modelling wildfire occurrence: the case of Catalonia. *Environmental and Ecological Statistics*, *21*, 531-563.

Strobl, C., Boulesteix, A., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, *9*, 307.

Strobl, C., Boulesteix, A., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, *8*, 25.

Variable	**Abbreviation**	**Description**	**Source**[4]
$V_{\sigma,1}$	House	The total number of houses	IFV
$V_{\sigma,2}$	House_indu	The number of houses with an industrial function	IFV
$V_{\sigma,3}$	House_hotl	The number of houses with a hotel function	IFV
$V_{\sigma,4}$	House_resi	The number of houses with a residential function	IFV
$V_{\sigma,5}$	House_20	The number of houses constructed before 1920	IFV
$V_{\sigma,6}$	House_2045	The number of houses constructed between 1920 and 1945	IFV
$V_{\sigma,7}$	House_4570	The number of houses constructed between 1945 and 1970	IFV
$V_{\sigma,8}$	House_7080	The number of houses constructed between 1970 and 1980	IFV
$V_{\sigma,9}$	House_8090	The number of houses constructed between 1980 and 1990	IFV
$V_{\sigma,10}$	House_90	The number of houses constructed after 1990	IFV
$V_{\sigma,11}$	House_frsd	The number of free standing (detached or semi-detached) houses	IFV
$V_{\sigma,12}$	House_nfrsd	The number of houses other than $V_{\sigma,11}$	IFV
$V_{\sigma,13}$	Resid	The number of residents	CBS
$V_{\sigma,14}$	Resid_14	The number of residents with an age in the range of 0 till 14	CBS
$V_{\sigma,15}$	Resid_1524	The number of residents with an age in the range of 15 till 24	CBS
$V_{\sigma,16}$	Resid_2544	The number of residents with an age in the range of 25 till 44	CBS
$V_{\sigma,17}$	Resid_4564	The number of residents with an age in the range of 45 till 64	CBS
$V_{\sigma,18}$	Resid_65	The number of residents with an age of 65 or higher	CBS
$V_{\sigma,19}$	Man	The number of male residents	CBS
$V_{\sigma,20}$	Woman	The number of female residents	CBS
$V_{\sigma,21}$	Address	The density of addresses in the neighbourhood	CBS
$V_{\sigma,22}$	Urbanity	The urbanity of the neighbourhood	CBS
$V_{\sigma,23}$	Town	Boolean variable indicating the presence of a town	CBS
$V_{\tau,1}$	WindSpeed	Daily mean wind speed (km/h)	KNMI
$V_{\tau,2}$	Temperature	Daily mean temperature (degree Celsius)	KNMI
$V_{\tau,3}$	WindChill	Daily mean wind chill (kg*cal/m2/h) (calculated from $V_{\tau,1}, V_{\tau,2}$)	
$V_{\tau,4}$	Sunshine	Daily sunshine duration calculated from global radiation (h)	KNMI
$V_{\tau,5}$	Visibility	Categorical variable indicating the daily minimum visibility[5]	KNMI
$V_{\tau,6}$	Season	Categorical variable indicating the season that a day belongs to	

Table 1: Environmental variables with their abbreviations, descriptions and sources.

[4]IFV: *Instituut Fysieke Veiligheid*, CBS: *Centraal Bureau voor de Statistiek*, KNMI: *Koninklijk Nederlands Meteorologisch Instituut*. All data is provided by the Twente Fire Brigade especially for this research.

[5]The variable, visibility, is well defined by weather technicians on certain rules, where minimum visibility distances (0–∞ *km*) are mapped to 1-80 levels. Hence, we dealt with it as a numerical variable in the random forest.