Data-driven chimney fire risk prediction using machine learning and point process tools

Marie-Colette van Lieshout

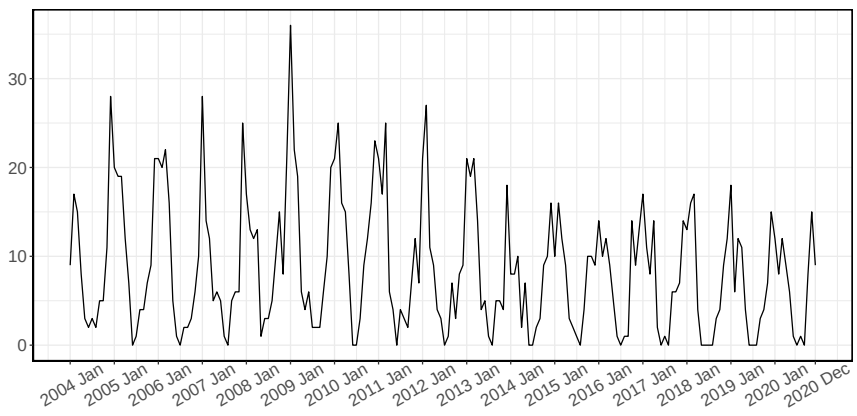
CWI & University of Twente
The Netherlands

Joint work with Changqing Lu, Maurits de Graaf and Paul Visscher.

May 2023

UNIVERSITY
OF TWENTE.

Chimney fire data



Chimney fires in Twente, 2004–2020.

Explanatory variables

NIPV, CBS, KNMI maintain information on population classified

- according to gender,
- age,
- address density, urbanity, ...,

on number of houses classified

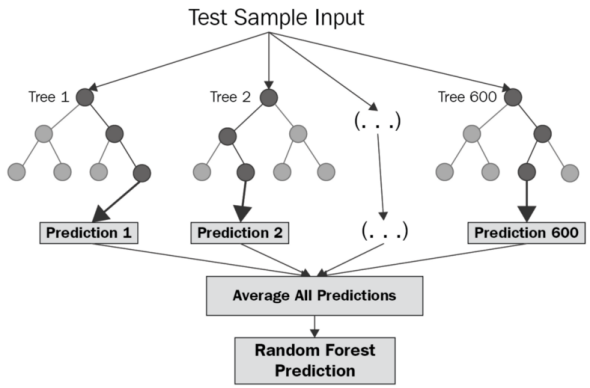
- according to function (industrial, residential, ...),
- date of construction,
- type (tower block, terrace, detached, semi-detached, ...)

and weather

- wind speed, temperature, sunshine, visibility, wind chill.

Random forests

x_i^j: putative explanatory variables, y_i: response.



Idea: split each node in two (e.g. $L = \{x_i^j \leq c\}$ and $R = \{x_i^j > c\}$) by minimising the residual sum of squares

$$\sum_{(y_i,(x_i^j)_j)\in L} (y_i - \bar{y}_L)^2 + \sum_{(y_i,(x_i^j)_j)\in R} (y_i - \bar{y}_R)^2.$$

Random forests - ctd

For root node, use bagging:

- bootstrap sample (size of data (n)) **with replacement** from $(y_i, (x_i^j)_j)$,
- reduces variance, and
- out-of-bag sample can be used to define errors.

At each node:

- select **mtry** (e.g. a third of) explanatory variables,
- find the best split based on the selected variables.

Allowing sub-optimal splits

- varies the trees, reduces bias and avoids local optima, but
- the resulting model may be hard to interpret.

Variable selection

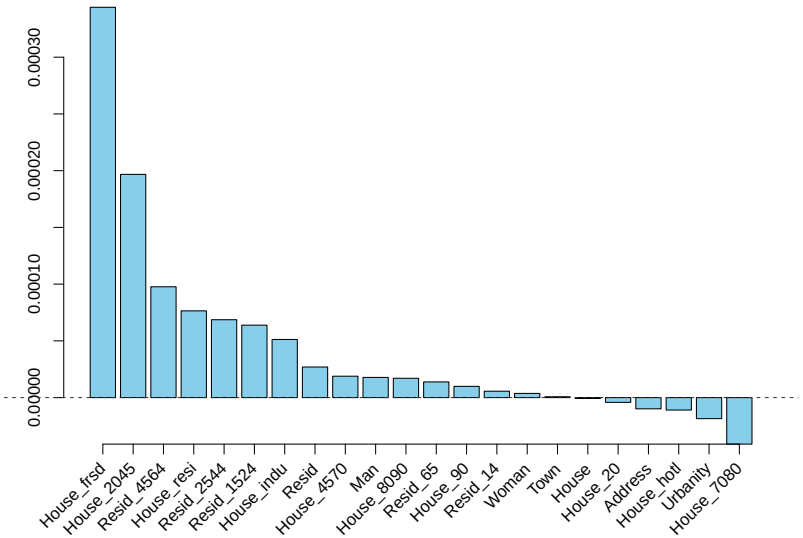
Consider explanatory variable x^j.
For each tree t, look at the out-of-bag data $(y_i, (x_i^j)_{j=1}^p)$ (indices denoted oB_t) and randomly permute the values of x_i^j to obtain $(y_i, (x_i^1, \ldots, x_i^{j-1}, x_{\pi_j(i)}^j, x_i^{j+1}, \ldots, x_i^p))$. Calculate

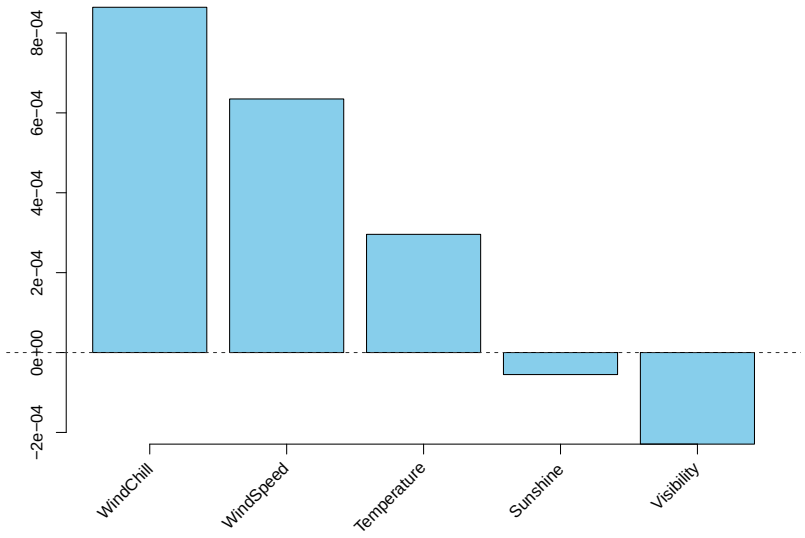
$$\boxed{I^t(x_j) = \frac{\sum_{i \in oB(t)} (y_i - \hat{y}_{i,\pi_j}^t)^2}{|oB(t)|} - \frac{\sum_{i \in oB(t)} (y_i - \hat{y}_i^t)^2}{|oB(t)|}.}$$

A big increase in mean squared error $I(x_j) = \text{mean}(I^t(x_j))$ means that explanatory variable x^j is important.

To avoid overestimating the importance of correlated variables, condition on the value of some variables (R-package party, Hothorn, Hornik, Strobl and Zeileis).

Selection of explanatory variables





Conclusion: use

- the number of freestanding houses and those build before the war,
- wind chill and wind speed.

Model assumptions

- chimneys catch fire independently,
- the rate is type-dependent,
- seasonal,
- subject to wind chill / speed

lead to a Poisson point process with intensity function

$$\boxed{\lambda(u,t) = \sum_{k=1}^{4} h_k(u)\,\lambda_k(t)}$$

where $h_k(u)$ is the density of houses of type k at location u and, writing C_1, C_2 for wind chill / speed,

$$\begin{aligned} \log \lambda_k(t) &= \text{Harmonic}(t) + \text{Polynom}(C_1(t)) \\ &+ \text{Polynom}(C_2(t)) + \text{Polynom}(C_1(t)C_2(t)) \end{aligned}$$

captures seasonal and weather fluctuation for type k at time t.

Logistic regression estimator

Find $\hat{\theta}$ that solves

$$\boxed{\sum_{x \in \mathbf{x}} \frac{\rho(x)/\lambda(x;\theta)}{\lambda(x;\theta)+\rho(x)} \nabla\lambda(x;\theta) - \sum_{d \in \mathbf{d}} \frac{1}{\lambda(d;\theta)+\rho(d)} \nabla\lambda(d;\theta) = 0}$$

(Baddeley et al, 2014).

We fit each house type separately and use

$$\rho_k(u,t) = r_k h_k(u) \left(\frac{1}{2} + \frac{1}{4} \left[\sin\left(\frac{2\pi}{365} t + \frac{pi}{2} \right) + 1 \right] \right).$$

Asymptotic framework

(C1) $\{Y_i\}$ and $\{E_i\}$, $i \in \mathbb{N}^+$, are independent sequences of i.i.d. point processes on some bounded open set $W \times T \subset \mathbb{R}^2 \times \mathbb{R}$,

$$X_n = \cup_{i=1}^n Y_i \text{ and } D_n = \cup_{i=1}^n E_i.$$

(C2) Y_i is a Poisson point process with intensity function

$$\lambda(u;\theta) = b(u) \exp\left[\theta^\top C(u)\right]$$

for integrable $b > 0$, measurable vector C of covariates and parameter vector $\theta \in \Theta \subset \mathbb{R}^m$, open.
E_i has integrable intensity function $\rho > 0$.

(C3) E_i has bounded pcf g.

Technical conditions

(C4) For every $\theta \in \Theta$ there exist $\epsilon_1(\theta)$, $\epsilon_2(\theta) > 0$ such that $\epsilon_1(\theta) < \inf_{u \in W \times T} \rho(u)/\lambda(u;\theta)$ and $\sup_{u \in W \times T} \rho(u)/\lambda(u;\theta) < \epsilon_2(\theta)$.

(C5) $\sup_{u \in W \times T} \|C(u)\| < \infty$.

(C6) Θ is convex.

(C7) $\lambda(u;\theta) = \lambda(u;\widetilde{\theta})$ almost everywhere on $W \times T$ implies $\theta = \widetilde{\theta}$.

(C8) The $[m \times m]$-dimensional matrix U, whose (k,l)-th entry reads

$$\int_{W \times T} \frac{\lambda(u;\theta_0)\rho(u)C_k(u)C_l(u)}{\lambda(u;\theta_0) + \rho(u)} \mathrm{d}u$$

with $\theta_0 \in \Theta$ is positive definite.

Remark: U (and V defined later) can be estimated consistently.

Strong consistency

Theorem

Assume that (C1)–(C2) and (C4)–(C7) hold. Define

$$l_n(\theta) = \sum_{x \in X_n} \log \left[\frac{\lambda(x;\theta)}{\lambda(x;\theta) + \rho(x)} \right] + \sum_{x \in D_n} \log \left[\frac{\rho(x)}{\lambda(x;\theta) + \rho(x)} \right], \quad \theta \in \Theta,$$

and set $\widehat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} l_n(\theta)$.

If $\widehat{\theta}_n$ *is attained, as* $n \to \infty$, $\widehat{\theta}_n$ *converges* P_{θ_0}*-almost surely to* θ_0.

Central limit theorem

Theorem

Assume that (C1)–(C8) hold. Define

$$s_n(\theta) = \sum_{x \in X_n} \frac{\rho(x)}{\lambda(x;\theta) + \rho(x)} C(x) - \sum_{x \in D_n} \frac{\lambda(x;\theta)}{\lambda(x;\theta) + \rho(x)} C(x), \quad \theta \in \Theta,$$

and define $\widehat{\theta}_n$ *by* $s_n(\widehat{\theta}_n) = 0$. *If* $\widehat{\theta}_n$ *is attained, as* $n \to \infty$,

$$n^{1/2}(\widehat{\theta}_n - \theta_0) \to^{P_{\theta_0}} \mathcal{N}(0, U^{-1} V (U^{-1})^\top)$$

where V *is an* $[m \times m]$*-dimensional matrix whose* (k, l)*-th entry reads*

$$\int_{W \times T} \frac{\lambda(u;\theta_0)\rho(u)C_k(u)C_l(u)}{\lambda(u;\theta_0) + \rho(u)} \mathrm{d}u +$$

$$\int_{(W \times T)^2} \frac{\lambda(u;\theta_0)\rho(u)C_k(u)}{\lambda(u;\theta_0) + \rho(u)} \frac{\lambda(v;\theta_0)\rho(v)C_l(v)}{\lambda(v;\theta_0) + \rho(v)} (g(u,v) - 1) \, \mathrm{d}u \mathrm{d}v.$$

Existence

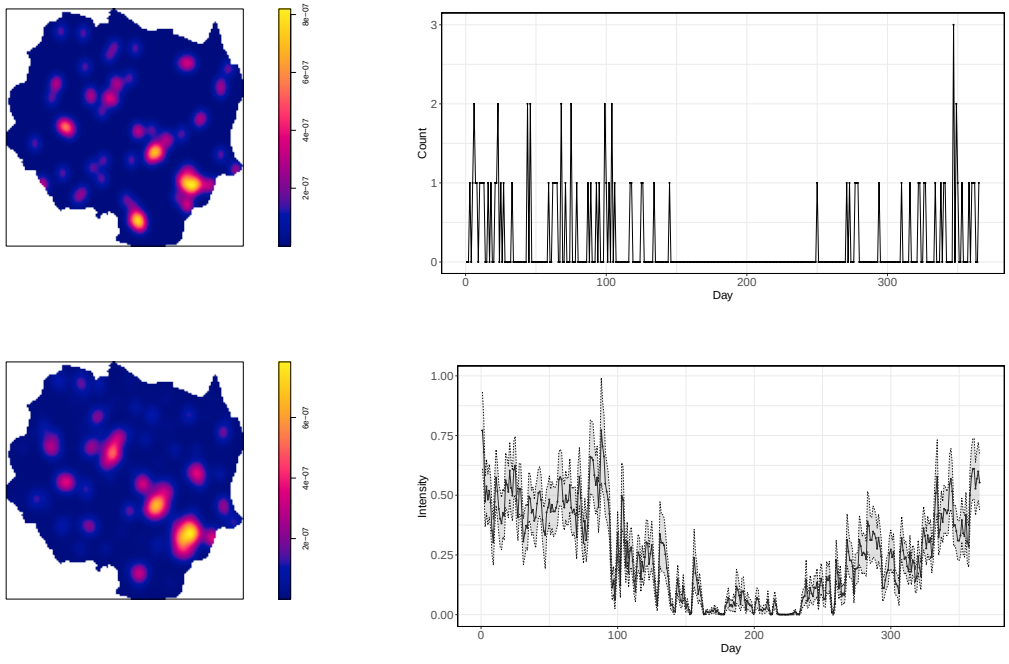
Theorem

Assume that (C1)–(C8) hold. Define

$$s_n(\theta) = \sum_{x \in X_n} \frac{\rho(x)}{\lambda(x;\theta) + \rho(x)} C(x) - \sum_{x \in D_n} \frac{\lambda(x;\theta)}{\lambda(x;\theta) + \rho(x)} C(x), \quad \theta \in \Theta.$$

Then, an estimator $\widehat{\theta}_n$ exists that solves $s_n(\widehat{\theta}_n) = 0$ with a probability tending to one as $n \to \infty$.

Back to chimney fires – predicted hazard for 2020



Top: Observed (smoothed) counts. Bottom: predicted fire hazard.
Units in space (left) per square meter and in time (right) per day.

Combination of machine learning and statistics

Random forests

- can detect both linear and non-linear correlation between a putative explanatory variable and chimney fire occurrences,
- nonparametrically,
- for a large number of candidates, and
- with conditional permutation importance techniques, the bias towards correlated variables is suppressed.

Statistical models

- rely on natural assumptions,
- are easier to interpret, and
- allow significance testing, model selection, and
- uncertainty quantification.

Further conclusions

- By tuning the dummy process,
  - the bias caused by 'saturation' (linear dependence on number of houses breaks down in large cities) is reduced, and
  - the fit in weather tipping months is improved.
- The asymptotic results can be extended to
  - non-Poisson models under replication, and
  - other estimating equations.

Further reading

C. Lu, M.N.M. van Lieshout, M. de Graaf and P.J. Visscher.
Data-driven chimney fire risk prediction using machine learning and point process tools.
Annals of Applied Statistics, to appear.

M.N.M. van Lieshout and C. Lu.
Infill asymptotics for logistic regression estimators for spatio-temporal point processes.
ArXiv 2208.12080, August 2022.