Predictive Inference of a Wildfire Risk Pipeline in the United States *Proposal Track*

Niccolò Dalmasso[⋆]**, Alex Reinhart**[⋆]**, Shamindra Shrotriya**[⋆]
Department of Statistics & Data Science
Carnegie Mellon University
Pittsburgh, PA 15213
sshrotri@stat.cmu.edu

1 Introduction

Wildfires are rare catastrophic events that are influenced by global climate change and present ongoing threats to life and property. The August 2019 IPCC report on climate change [1] notes that climate change is "expected to enhance the risk and severity of wildfires" in many areas. Hence there is an urgent need to accurately statistically model wildfire risks. Wildfire risk modeling requires accounting for several separate but related risk components [2] which can be viewed as a "wildfire risk pipeline". First, one must model where fires are most likely to start, based on weather, human activity, and related factors; then, based on a fire's location, model the fire's duration and size. Finally, a model can project risk exposure, the number of lives or properties exposed to the fire.

Prior work has generally modeled each component in isolation and typically limited to small regions, due to the varied spatiotemporal resolution and quality of open data available for modeling on a national scale. Both physical and stochastic models have been used to model the various wildfire risk components (for in-depth surveys see [2, 3]). Fire locations have typically been modeled as point process models [4, 5], fit by maximum likelihood over discretized space-time grids [6]. Logistic Generalized Additive Models (GAMs) have been used to model seasonal non-linear relationships among fire occurrence and covariates [7]. Fire duration is usually modeled via survival analysis techniques [8]; as duration is typically heavy-tailed, the baseline survivor functions are modeled as Gaussian, Gumbel or logistic distributions [2]. Parametric heavy tailed distributions are often used for fire size as well (e.g. tapered Pareto [9], Generalized Extreme Value (GEV) distribution [10] and generalized Pareto distribution with additional environmental features as inputs [11]). Previous work has generally sought to model each component of wildfire risk separately. More specifically, fire occurrence was combined with an independent survival model [12, 13] and bivariate extreme value models were used in marked point process settings to explicitly model dependence between wildfire risk components [14, 15].

Our contribution in this paper is threefold. First, we seek to provide end-to-end modeling of the wildfire risk pipeline with an emphasis on both predictive accuracy and *uncertainty* for each risk estimate in the pipeline. Our proposed model accounts for fire location, size, duration, and risk exposure sequentially, so that uncertainty in each step can be propagated to later steps. Second, we seek to build our models using on the entire continental United States using open data, rather than limiting our analysis to a specific state or county. Third, we provide open-source code [1] to download, transform and aggregate open data relevant to wildfire prediction in the continental United States. We hope this will set an openly available national benchmark for wildfire risk modeling.

[1] https://github.com/shamindras/backburner.

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

[⋆] Authors contributed equally.

2 Data Aggregation Pipeline

Our first goal is to provide open-source code to *extract*, *transform*, and *load* (ETL) publicly available wildfire-related data in the continental United States to produce a single database containing all information relevant to wildfires. The first release includes the following data sources: **(i)** Wildfire perimeters from both the Monitoring Trends in Burn Severity project (MTBS, [16], 1984–2016) and Geospatial Multi-Agency Coordination (GeoMAC, [17], 2000–2019) **(ii)** Weather data from the National Oceanic and Atmospheric Administration (NOAA), specifically the daily global historical climatology network and storm events database ([18, 19]) **(iii)** Wildfire data from the US Forest Service archive ([20], up to 2015) and **(iv)** Lightning strikes from the National Lightning Detection Network (NDLN, [21], 1986–2018). Our code conveniently consolidates all these disparate data sources into a single open geospatial SQL database.

3 Wildfire Prediction Model

We propose to model wildfire occurrence as a spatiotemporal point process. Each observed wildfire i is an event (s_i, t_i) comprising the fire's 2D spatial location $s_i \in X \subset \mathbb{R}^2$, and $t_i \in T \subset \mathbb{R}$, the time the fire occurred. Each fire also has additional features, known as *marks*: the fire's duration d_i, fire size z_i, and the exposure risk of the fire c_i (e.g. lives at risk). The point process model has several parts. The *ground process* $\lambda_g(s, t)$ models the rate of wildfires per unit time per unit space, and varies according to features of the location, season, weather, and so on [22]. It can be defined as a parametric function of spatial covariates or as a nonparametric model. The distributions of d_i, z_i, and c_i also vary with these covariates, and by location and time, so they are modeled with conditional densities $f_D(d \mid s, t)$, $f_Z(z \mid s, t, d)$, and $f_C(c \mid d, s, t)$. The overall model is

$$\lambda(s, t, z, d, c) = \lambda_g(s, t) f_D(d \mid s, t) f_Z(z \mid s, t, d) f_C(c \mid s, t, d, z), \tag{1}$$

with the log-likelihood function

$$\ell(\theta) = \sum_{i=1}^{n} \log(\lambda(s_i, t_i, d_i, z_i, c_i)) - \int_T \int_X \lambda_g(s, t)\, \mathrm{d}s\, \mathrm{d}t,$$

where θ is the vector of model parameters, i indexes observed fires, and X and T define the space and time for which fires were observed [23]. The model can be fit using maximum likelihood [24]. The power of this model is that it can be easily interpreted in various ways and key quantities can be obtained straightforwardly. For instance, the expected number of fires in a spatial region S and temporal window T is the integral of the ground process over that spatio-temporal window $\mathbb{E}[N(S \times T)] = \int_S \int_T \lambda_g(s, t)\, \mathrm{d}t\, \mathrm{d}s$, where N is the counting measure. In the same fashion, to calculate the number of fires expected of particular sizes or costs, the intensity $\lambda(s, t, z, d, c)$ can be integrated over these sizes or costs as well. In general, the factorization in Eq. 1 also allows for predictive inference at different part of the pipeline: calculating statistics of interest for duration can be done by integrating $f_D(d \mid s, t)$ only over a spatial region and temporal window, hence not requiring the full pipeline to be run. The key statistical quantities of interest and modeling output at each stage of the pipeline are summarized in Table 1.

	Stage 1 **Fire Occurrence**	**Stage 2** **Fire Duration**	**Stage 3** **Fire Size**	**Stage 4** **Risk Exposure**
Quantity of Interest	$\lambda_g(x, y, t)$	$f_D(d \mid s, t)$	$f_Z(z \mid s, t, d)$	$f_C(c \mid s, t, d, z)$
Modeling Method	MLE	CDE	CDE	CDE
Modeling Output	$\mathbb{E}[N(S \times T)]$	Duration Density	Size Density	Risk Density

Table 1: Proposed model pipeline for wildfire locations and risks

We propose to model the conditional densities at each stage using conditional density estimation (CDE) techniques. This can be done via fitting suitable parametric family models such as heavy tailed distributions [9, 10], where the distribution are chosen based on the domain knowledge. Another approach is to estimate conditional densities nonparametrically, for instance relying on nonparametric regression method such as nearest neighbor, random forest and kernel density estimate [25, 26, 27, 28] or by making assumptions on the form of the conditional distribution [29, 30]. The key advantage of using conditional densities, rather than simple regression models, is they make uncertainty quantification in prediction straightforward, as the full conditional distribution is available for simulation. This also allows to propagate uncertainty through the pipeline in a forward fashion; for instance, given the spatiotemporal coordinates of a wildfire s, t, one can sample its duration from $f_D(d|s, t)$, then its size from $f_Z(z|s, t, d)$ and finally its risk exposure $f_C(c|s, t, d, z)$. Repeating this process multiple times can provide uncertainty over key quantities of interest of a wildfire, at any stage of the pipeline. As conditional density estimation techniques are negatively affected by small or skewed training data, we plan to validate the fit at each stage of the pipeline using e.g., probability integral transforms and highest predictive density regions [31, 25, 30]. We also intend to assess goodness of fit and validate predictive accuracy at each stage of the model pipeline. Spatiotemporal point process residual diagnostic techniques are surveyed in detail in [32] including using Voronoi residual maps [33].

References

[1] Intergovernmental Panel on Climate Change. *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems*. April 2019.

[2] Dexen D.Z. Xi, Stephen W. Taylor, Douglas G. Woolford, and C.B. Dean. Statistical models of key components of wildfire risk. *Annual Review of Statistics and Its Application*, 6(1):197–222, 2019.

[3] S. W. Taylor, Douglas G. Woolford, C. B. Dean, and David L. Martell. Wildfire prediction to inform fire management: Statistical science challenges. *Statistical Science*, 28(4):586–615, 2013.

[4] Yonghe Wang and Kerry R Anderson. An evaluation of spatial and temporal patterns of lightning-and human-caused forest fires in Alberta, Canada, 1980–2007. *International Journal of Wildland Fire*, 19(8):1059–1072, 2011.

[5] Justin Podur, David L. Martell, and Ferenc Csillag. Spatial point pattern analysis of lightning-caused forest fires in the boreal forest region of Ontario. In G.J. Arthaud and T.M Barrett, editors, *Systems Analysis in Forest Resources*, Managing Forest Ecosystems, pages 61–68. Springer, 2003.

[6] David R Brillinger, Haiganoush K Preisler, and John W Benoit. Risk assessment: a forest fire example. In Darlene R. Goldstein, editor, *Statistics and science: a Festschrift for Terry Speed*, pages 177–196. Institute of Mathematical Statistics, 2003.

[7] Haiganoush K. Preisler and Alan A. Ager. Forest-fire models. In *Encyclopedia of Environmetrics*. John Wiley & Sons, Ltd, 2013.

[8] Amy A. Morin, Alisha Albert-Green, Douglas G. Woolford, and David L. Martell. The use of survival analysis methods to model the control time of forest fires in Ontario, Canada. *International Journal of Wildland Fire*, 24(7):964, 2015.

[9] Frederic Paik Schoenberg, Roger Peng, and James Woods. On the distribution of wildfire sizes. *Environmetrics*, 14(6):583–592, 2003.

[10] Enrique Castillo. *Extreme value theory in engineering*. Academic Press, 2012.

[11] A.C. Davison and R. Huser. Statistics of extremes. *Annual Review of Statistics and Its Application*, 2(1):203–235, 2015.

[12] Amy A Morin. A spatial analysis of forest fire survival and a marked cluster process for simulating fire load. Master's thesis, The University of Western Ontario, 2014.

[13] Haiganoush K. Preisler, Anthony L. Westerling, Krista M. Gebert, Francisco Munoz-Arriola, and Thomas P. Holmes. Spatially explicit forecasts of large wildland fire probability and suppression costs for California. *International Journal of Wildland Fire*, 20(4):508, 2011.

[14] Jonathan Yoder and Krista Gebert. An econometric model for ex ante prediction of wildfire suppression costs. *Journal of Forest Economics*, 18(1):76–89, 2012.

[15] Jude Bayham. *Characterizing incentives: an investigation of wildfire response and environmental entry policy*. PhD thesis, Washington State University, 2013.

[16] Monitoring Trends in Burn Severity (MTBS). Wildfire perimeters. data retrieved from `https://rmgsc.cr.usgs.gov/outgoing/GeoMAC/historic_fire_data/`.

[17] Geospatial Multi-Agency Coordination (GeoMAC). Wildfire perimeters. data retrieved from `https://rmgsc.cr.usgs.gov/outgoing/GeoMAC/historic_fire_data/`.

[18] Global Historical Climatology Network (GHCN). Climate summaries from land surface stations across the globe. data retrieved from `ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/by_year/`.

[19] Storm Events Database. Occurrence of storms and significant weather phenomena. data retrieved from `ftp://ftp.ncdc.noaa.gov/pub/data/swdi/stormevents/csvfiles/`.

[20] Karen C Short. Spatial wildfire occurrence data for the United States, 1992–2015. *Forest Service Research Data Archive*, 2017.

[21] National Oceanic and Atmospheric Administration. NOAA severe weather data inventory. `https://www.kaggle.com/noaa/noaa-severe-weather-data-inventory`.

[22] Roger D Peng, Frederic P Schoenberg, and James Woods. Multi-dimensional point process models for evaluating a wildfire hazard index. Unpublished manuscript, `https://escholarship.org/uc/item/4r37990g`, 2003.

[23] D J Daley and D Vere-Jones. *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Springer, 2nd edition, 2003.

[24] Haiyong Xu and Frederic Paik Schoenberg. Point process modeling of wildfire hazard in Los Angeles County, California. *The Annals of Applied Statistics*, 5(2A):684–704, 2011.

[25] Rafael Izbicki, Ann B Lee, and Peter E Freeman. Photo-z estimation: An example of nonparametric conditional density estimation under selection bias. *The Annals of Applied Statistics*, 11(2):698–724, 2017.

[26] Peter E Freeman, Rafael Izbicki, and Ann B Lee. A unified framework for constructing, tuning and assessing photometric redshift density estimates in a selection bias setting. *Monthly Notices of the Royal Astronomical Society*, 468(4):4556–4565, 2017.

[27] Taylor Pospisil and Ann B. Lee. RFCDE: Random Forests for Conditional Density Estimation. *arXiv e-prints*, page arXiv:1804.05753, Apr 2018.

[28] Taylor Pospisil and Ann B. Lee. (f)RFCDE: Random Forests for Conditional Density Estimation and Functional Data. *arXiv e-prints*, page arXiv:1906.07177, Jun 2019.

[29] Rafael Izbicki and Ann B Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831, 2017.

[30] Niccolò Dalmasso, Taylor Pospisil, Ann B. Lee, Rafael Izbicki, Peter E. Freeman, and Alex I. Malz. Conditional density estimation tools in Python and R with applications to photometric redshifts and likelihood-free cosmological inference. arXiv preprint 1908.11523, 2019.

[31] Kai Lars Polsterer, Antonio D'Isanto, and Fabian Gieseke. Uncertain photometric redshifts. arXiv preprint arXiv:1608.08016, 2016.

[32] Alex Reinhart. A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33(3):299–318, 2018.

[33] Andrew Bray, Ka Wong, Christopher D. Barr, and Frederic Paik Schoenberg. Voronoi residual analysis of spatial point process models with applications to California earthquake forecasts. *Ann. Appl. Stat.*, 8(4):2247–2267, 12 2014.