Extreme Value Analysis 2021 Conference: Data Competition

Thomas Opitz, Biostatistics and Spatial Processes, INRAE, France

Co-organization and logistic support: Raphaël Huser (KAUST), Ioannis Papastathopoulos (University of Edinburgh)

To register, send a message to thomas.opitz@inrae.fr after reading the instructions.

1 Wildfires and their extremes: a global challenge

Wildfires are uncontrolled fires of combustible material composed of natural vegetation, such as forests or shrubland. They represent an environmental hazard with major impacts worldwide, and their frequency of occurrence and size is expected to increase with global warming (Jones *et al.*, 2020). Each year, wildfires cause many direct human casualties, and they are at the origin of extreme air pollution episodes and the loss of biodiversity and other ecosystem services. They further contribute an important fraction of global greenhouse gas emissions each year, such that they can further exacerbate climate change. Wildfire modeling has been tackled with a wide variety of statistical approaches (see, *e.g.*, Preisler *et al.*, 2004; Pereira and Turkman, 2019; Xi *et al.*, 2019), some of them invoking extreme-value theory (see, *e.g.*, Pereira and Turkman, 2019).

As to the origin of wildfire ignitions, lightning is the principal natural cause, but in the majority of cases human activities are responsible. They may be intentional (arson, children) or accidental (debris burning, agricultural activities, campfires, smoking). Typically, wildfires are the result of the concurrence of the presence of combustible material (*e.g.*, forest), its easy flammability (*e.g.*, resulting from extreme weather conditions such as droughts), and a trigger (*e.g.*, lightning or human activity). In most wildfire-prone areas over the globe, wildfire activity shows seasonal cycles due to the seasonality of favorable weather conditions.

To aid in wildfire management, it is crucial to understand and predict the risk factors contributing to wildfire activity, and their spatio-temporal distribution. Wildfire management includes a multitude of tasks, including monitoring of forest ecosystems, deployment of preventive measures, firefighting logistics, and short-term forecasting and long-term projections of wildfire activity.

In this data competition, we focus on two important components of wildfire activity: wildfire occurrence, and wildfire size. Given a region of space and a period of time, we consider the number of spatially separated wildfire events as an observation with respect to the first aspect (occurrence), and the aggregated burnt area of wildfires originating in the area of interest as an observation with respect to the second aspect (size). In light of the relatively heavy tails in observations of both of these variables, combined with the lack of "smoothness" of such variables over space and time, their accurate prediction is a challenging task. The most extreme impacts and the biggest difficulties in wildfire management are associated to large values of burnt area and/or of wildfire counts, for both of which prediction is very difficult.

2 Dataset description

2.1 Structure of provided data table

A comprehensive wildfire dataset covering the period from 1993 to 2015 for the continental United States is used. We exclude the state of Alaska and islands such as Hawaiï. Spatial coordinates are systematically given in the WGS84 system, that is, the usual longitude and latitude coordinates. Based on a $0.5^o \times 0.5^o$ grid of longitude and latitude coordinates (roughly 55 by 55 km) covering the study area, the wildfire data to be used in this competition comprise the number of wildfires and the aggregated burnt area in each grid cell for each month (between March and September) of the observation period. Moreover, 35 auxiliary variables related to land cover, weather and altitude are provided at the same spatial and temporal resolution, and can be utilized for modeling.

The dataset is provided as an `RData` file containing an `R` dataframe called `data_train_DF` with the following columns:

1. CNT – number of wildfires (values to be predicted are given as `NA`)
2. BA – aggregated burnt area of wildfires in acres (values to be predicted are given as `NA`)
3. lon – longitude coordinate of grid cell center
4. lat – latitude coordinate of grid cell center
5. area – the proportion of a grid cell that overlaps the continental US (a value in $(0, 1]$, which can be smaller than 1 for grid cells on the boundary of the US territory)
6. month – month of observation (integer value between 3 and 9)
7. year – year of observation (integer value between 1 and 23, with 1 corresponding to year 1993 and 23 to year 2015)
8. lc_1 to lc_{18}: area proportion of 18 land cover classes in the grid cell (see details below)
9. altiMean, altiSD: altitude-related variables given as mean and standard devation in the grid cell (see details below)
10. clim_1 to clim_{10}: monthly means of 10 meteorological variables in the grid cell (see details below)

Note that the area proportions lc_1 to lc_{18} do not always sum to exactly 1 for each pixel and month since a few classes with quasi-0 proportion have been removed. These 18 predictors are therefore almost *collinear*, *i.e.*, $\text{lc}_{i_0} \approx 1 - \sum_{i=1,\ i\neq i_0}^{18} \text{lc}_i$ for $i_0 = 1, \ldots, 18$.

2.2 Original datasets and their preprocessing

The dataset provided for the competition has been composed using several data sources. This subsection provides some background.

2.2.1 US wildfires

We use a comprehensive dataset of wildfires in the US, which has been gathered from various wildfire inventories. It makes available a set of unified attributes available for each wildfire. In this data competition, we use the geographic position, the time of occurrence and the burnt area of individual wildfires. This information is aggregated towards the monthly longitude-latitude grid by counting the number of wildfires within each grid point (variable CNT), and by summing up the burnt areas of these wildfires (variable BA).

2.2.2 Land cover

The land monitoring service of the European Union's COPERNICUS service for remote sensing produces global land cover classification maps at 300 m spatial resolution and annual temporal resolution. Data are provided online through the COPERNICUS Climate Data Service. The classification uses 38 classes whose definition is based on the United Nations Food and Agriculture Organizations (UN FAO) Land Cover Classification System (LCCS). Of these 38 categories, only 18 are observed with nonnegligible proportion in the study area. For the data competition, data are aggregated to the $0.5^o \times 0.5^o$ grid of longitude and latitude by considering the proportion of each of the 18 categories within each grid cell. Proportions of the following categories are considered in lc_1 to lc_{18} in the provided data. Their denominations as given in the original dataset are as follows:

1. cropland rainfed
2. cropland rainfed herbaceous cover
3. mosaic cropland
4. mosaic natural vegetation
5. tree broadleaved deciduous closed to open
6. tree broadleaved deciduous closed
7. tree needleleave evergreen closed to open
8. tree needleleave evergreen closed
9. tree mixed
10. mosaic tree and shrub
11. shrubland
12. grassland
13. sparse vegetation
14. tree cover flooded fresh or brakish water
15. shrub or herbaceous cover flooded

16. urban
17. bare areas
18. water

2.2.3 Meteorological variables

The meteorological variables provided in the dataset of the competition are based on the gridded output of monthly means obtained within the ERA5-reanalysis on Land surface, available for a global grid of resolution $0.1^o \times 0.1^o$ from the COPERNICUS Climate Data Service. The following 10 variables (corresponding to clim_1 to clim_{10}, respectively, in the provided data) are considered for this data competition, with units given in parentheses:

1. 10m U-component of wind (the wind speed in Eastern direction) (m/s)
2. 10m V-component of wind (the wind speed in Northern direction) (m/s)
3. Dewpoint temperature (temperature at 2m from ground to which air must be cooled to become saturated with water vapor, such that condensation ensues) (Kelvin)
4. Temperature (at 2m from ground) (Kelvin)
5. Potential evaporation (the amount of evaporation of water that would take place if a sufficient source of water were available) (m)
6. Surface net solar radiation (net flux of shortwave radiation; mostly radiation coming from the sun) (J/m^2)
7. Surface net thermal radiation (net flux of longwave radiation; mostly radiation emitted by the surface) (J/m^2)
8. Surface pressure (Pa)
9. Evaporation (of water) (m)
10. Precipitation (m)

2.2.4 Variables related to altitude

Finally, two variables related to altitude are made available. The variable `altiMean` provides the mean altitude for each cell of the longitude-latitude grid, and `altiSD` provides the corresponding standard deviation. Original gridded data were provided by the Shuttle Radar Topography Mission (SRTM) at 90 m spatial resolution.

2.3 Training and validation datasets

The dataset has been split into a training dataset and a validation dataset. Validation is carried out based on the predictions for the variables of aggregated burnt areas (BA) and counts (CNT).

No data have been masked for uneven years (1993, 1995...,2015). For even years (1994, 1996, ..., 2014), overall $80,000$ observations of each of the two variables are used for validation; that is, they have been masked by setting them to `NA` in the dataset. The spatial and temporal positions of validation data are not completely random, but they tend to be clustered in space and time. Moreover, the validation locations for BA and CNT are not the same, but they are correlated. This means that the probability of having to validate both BA and CNT for a given grid cell and month is higher than the product of the two probabilities of having to validate BA or CNT.

3 Prediction goal, evaluation criterion and benchmark

3.1 Prediction goal

The aim of prediction is to estimate predictive distributions for the BA and CNT validation data, i.e., the data masked in the training dataset. More precisely, the value of the predictive distribution function of each `NA` observation of BA and CNT must be estimated for a list of severity thresholds.

For CNT, 28 severity thresholds are fixed as follows:

$$\mathcal{U}_{CNT} = \{u_1, u_2, \ldots, u_{28}\} = \{0, 1, 2, \ldots, 9, 10, 12, 14, \ldots, 30, 40, 50, \ldots, 100\}.$$

For each validation data point i and each $u \in \mathcal{U}_{CNT}$, the probability $p_{CNT,i}(u) = \mathbb{P}(\mathrm{CNT}_i \leq u)$ must be estimated. For BA, the following 28 severity thresholds are applied:

$$\begin{aligned}\mathcal{U}_{BA} = \{u_1, u_2, \ldots, u_{28}\} = \{&0, 1, 10, 20, 30, \ldots, 100, 150, 200, 250, 300, 400, 500, 1000,\\ &1500, 2000, 5000, 10000, 20000, 30000, 40000, 50000, 100000\}.\end{aligned}$$

For each validation data point i and each $u \in \mathcal{U}_{BA}$, the probability $p_{BA,i}(u) = \mathbb{P}(\mathrm{BA}_i \leq u)$ must be estimated.

3.2 Prediction scores

The quality of predictions will be compared among teams using a prediction score, and teams will be ranked based on the prediction score, with lower scores resulting in better rank. Separate prediction scores will be calculated for BA and CNT, resulting in separate rankings of teams participating in one or both of the sub-competitions of predicting BA and CNT, respectively. The overall prediction score, used to determine the overall winning team, results from adding up these two separate scores and ranking them. **A relatively lower score is better and will lead to a better ranking of the corresponding team.**

The scores used for the competition are variants of weighted ranked probability scores, which put relatively strong weight on good prediction in the extremes of the distribution of counts and burnt areas.

We denote by $\widehat{p}_{CNT,i}(u)$ the predicted probability for $\mathbb{P}(\text{CNT}_i \leq u)$ and by $\widehat{p}_{BA,i}(u)$ the predicted probability for $\mathbb{P}(\text{BA}_i \leq u)$. There are $k_{CNT} = k_{BA} = 80,000$ values to be predicted for CNT and BA, respectively. For the counts in CNT to be predicted, we use the score

$$S_{CNT} = \sum_{i=1}^{k_{CNT}} \sum_{u \in \mathcal{U}_{CNT}} \omega_{CNT}(u) \left(\mathbb{I}((u \leq \text{CNT}_i) - \widehat{p}_{CNT,i}(u)\right)^2.$$

where the weight function is given as

$$\omega_{CNT}(u) = \frac{\tilde{\omega}_{CNT}(u)}{\tilde{\omega}_{CNT}(u_{28})}, \quad \tilde{\omega}_{CNT}(u) = 1 - (1 + (u+1)^2/1000)^{-1/4},$$

and $\mathbb{I}$ is the indicator function defined as

$$\mathbb{I}(u \leq x) = \begin{cases} 1, & \text{if } u \leq x \\ 0, & \text{if } u > x. \end{cases}$$

The division by $\tilde{\omega}_{CNT}(u_{28})$ above ensures that the largest weight $\omega_{CNT}(u_{28})$ is 1. By analogy, the score for burnt areas in BA is

$$S_{CNT} = \sum_{i=1}^{k_{BA}} \sum_{u \in \mathcal{U}_{BA}} \omega_{BA}(u) \left(\mathbb{I}((u \leq \text{BA}_i) - \widehat{p}_{BA,i}(u)\right)^2$$

where

$$\omega_{BA}(u) = \frac{\tilde{\omega}_{BA}(u)}{\tilde{\omega}_{BA}(u_{28})}, \quad \tilde{\omega}_{BA}(u) = 1 - (1 + (u+1)/1000)^{-1/4}.$$

The overall score is $S_{TOTAL} = S_{CNT} + S_{BA}$.

A benchmark score is provided. For CNT, it corresponds to a generalized linear model with Poisson response distribution and log-link using all available covariates. For positive values of BA, it corresponds to fitting a generalized linear model with Gaussian response and log-link using all available covariates. The probability predictions of BA are obtained by combining the log-Gaussian BA model (for BA>0) with the probability of CNT$=0$ obtained from the Poisson model.

4 Instructions to submit predictions

Each team has to provide two matrices (of class "`matrix`" in `R`) with the following properties:

1. The matrices must be named `prediction_ba` and `prediction_cnt`, and saved into an `R` object called `prediction_teamname.RData` by using the `R` function `save(...)`. A code example of how to save the matrices in `R` for a team named FIREFIGHTERS is as follows:

   ```
   save(prediction_ba, prediction_cnt, file = "prediction_FIREFIGHTERS.RData")
   ```

2. The matrices `prediction_ba` and `prediction_cnt` must be of dimension $80,000 \times 28$, where the (i,j)-th entry corrresponds to the prediction $\widehat{p}_{BA,i}(u_j)$ or $\widehat{p}_{CNT,i}(u_j)$, respectively. The first line ($i = 1$) corresponds to the first `NA` value of BA or CNT, respectively, in the data matrix `data_train_DF` (starting at line 1), the second line ($i = 2$) to the second `NA` value of BA or CNT, respectively, encountered in the data matrix, and so on.

In addition, each team has to provide a clean and commented code to be able to reproduce the results if needed.

Each team will be provided with a link to a Dropbox folder to submit the results and the code.

5 Timetable and deadlines

1. Preliminary prediction 1 (optional): 28 February 2021 at 23:59 UTC
2. Preliminary prediction 2 (optional): 30 April 2021 at 23:59 UTC
3. **Final prediction: 31 May 2021 at 23:59 UTC**
4. EVA 2021 Conference: 28 June–02 July 2021

Each team can submit up to two preliminary predictions to compare their approach to the predictions of other teams. The final ranking will be based solely on the final prediction. Preliminary and final rankings (along with prediction scores) will be published on the conference web page.

6 Rules

1. There is no limit to the number of teams or team members. However, each participant can only be part of one team, not of several teams.
2. Only the final submission will be taken into account to rank the teams.
3. Submission of preliminary predictions is not mandatory, but highly encouraged.
4. Results must be submitted as specified above in §4, and the submitted `R` code must be clean and properly commented to be able to reproduce the results if needed.
5. Reverse engineering, or the use of other data sources (covariates not included in the provided dataset etc.), is strictly prohibited.
6. Late submissions will not be considered.
7. Failure to comply with the above rules may result in disqualification.

7 Rewards

1. The rankings will be published on the EVA 2021 website and in the *Extremes* journal. Teams can choose not to appear in the published rankings. The winners will be officially announced during the EVA 2021 conference.
2. The best-ranked teams will be invited to present their work during an invited session at the (online) EVA 2021 conference, organized by members of the School of Mathematics of the University of Edinburgh during the week of 28 June–02 July 2021.

3. After the EVA 2021 conference, all the teams will be invited to submit a paper describing their approach for publication in the journal *Extremes*. The submitted papers will undergo the usual peer-review process with the same quality standards and criteria of acceptance.

8 Getting Started

1. **Registration:** Register your team (and specify a team name) by sending an email to Thomas Opitz (thomas.opitz@inrae.fr). You will then receive a link to the data repository, an `R` script to load the data, and instructions on how to submit your predictions.

2. Open `R` on a terminal or using `RStudio`, open the script `Competition.R`, and follow the instructions.

References

Jones, M. W., Smith, A., Betts, R., Canadell, J. G., Prentice, I. C. and Le Quéré, C. (2020) Climate change increases risk of wildfires. *ScienceBrief Review* .

Pereira, J. and Turkman, K. (2019) Statistical models of vegetation fires: Spatial and temporal patterns. In *Handbook of Environmental and Ecological Statistics*, pp. 401–420. Chapman and Hall/CRC.

Preisler, H. K., Brillinger, D. R., Burgan, R. E. and Benoit, J. (2004) Probability based models for estimation of wildfire risk. *International Journal of wildland fire* **13**(2), 133–142.

Xi, D. D., Taylor, S. W., Woolford, D. G. and Dean, C. (2019) Statistical models of key components of wildfire risk. *Annual review of statistics and its application* **6**, 197–222.