École Polytechnique Fédérale de Lausanne

Report

Wildfires Data Challenge

Mongi Nouira

May 23, 2021

2 Exploratory Data Analysis

We focus on two important components of wildfire activity: wildfire occurrence and size. A comprehensive wildfire dataset covering the period from 1993 to 2015 for the continental United States is used. The states of Alaska and islands such as Hawaii are excluded. The goal of this challenge is to estimate the predictive distributions of the number of wildfires (CNT) and aggregated burnt area of wildfires in acres (BA) for a validation set. The scores used for the competition are variants of weighted ranked probability scores but to summarize the lower the score the better.

2.1 Data Structure

The dataset contains the following variables as columns :

- CNT : number of wildfires (values to be predicted are given as NA)
- BA : aggregated burnt area of wildfires in acres (values to be predicted are given as NA)
- lon : longitude coordinate of grid cell center
- lat : latitude coordinate of grid cell center
- 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)
- month : month of observation (integer value between 3 and 9)
- year : year of observation (integer value between 1 and 23, with 1 corresponding to year 1993 and 23 to year 2015)
- lc1 to lc18 : area proportion of 18 land cover classes in the grid cell
- altiMean, altiSD : altitude-related variables given as mean and standard deviation in the grid cell
- clim1 to clim10 : monthly means of 10 meteorological variables in the grid cell

We are provided with 563983 rows and 37 columns from which 111053 rows contain a missing value either CNT or BA or both. There are exactly 80000 missing values for the CNT and BA respectively. So one have multiple possibilities for training models :

First, to predict BA for example, we could use two models one using the available CNT for approximately 39% of the validation set and 61% without CNT. Second, we could also have a joint model for both response variables that could capture more properties such as the interaction between them. Finally, we could use only one model that predicts one response variable without using the other.

In real applications, the last two possibilities are the more appropriate to use for predictions since we don't know both variables future values. If we just want to understand the phenomenon of wildfires for non predictive risk management, then we could use the first possibility. For this challenge, since we want to have a better score, we should try the first idea.

2.2 Spatial-Temporal and Response Variables

Looking at the spacial distribution of the mean number of fires across the USA [Figure 1], we can see that some regions have more wildfires than the others such as California, Arizona, New Mexico, Georgia, South and North Carolina and New Jersey. For the aggregated burned areas, we have again California but also some states in the north west which don't have a high number of fires but have the biggest fire sizes. Also, the states on the south east seems to have a good control of fires since the size of burned area is rather small compared to the number of fires in those regions.

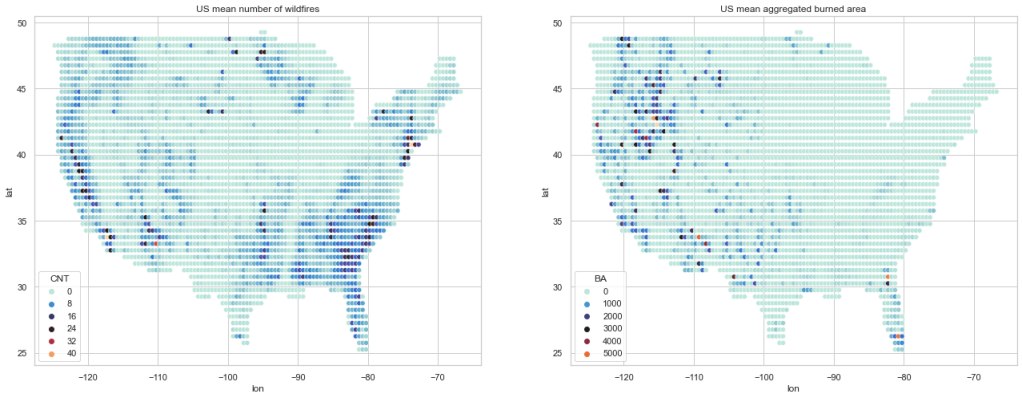

Figure 1: Spatial distribution of CNT and BA

We can also check if there is any correlation or a distribution pattern between the number of wildfires and the burnt areas [Figure 2]. We can see that a high number of fires don't necessarily mean a huge burned area and vice versa. Actually, most of the huge fires may have started once and continue for several days, that is why we see a peak for CNT = 1. Also looking at the log transformation of both variables for fires only, we may see a small increasing trend and a concentration of BA for high number of fires. Again, a small number of fires or even a single one can have very different values of BA which is represented as vertical straight lines.

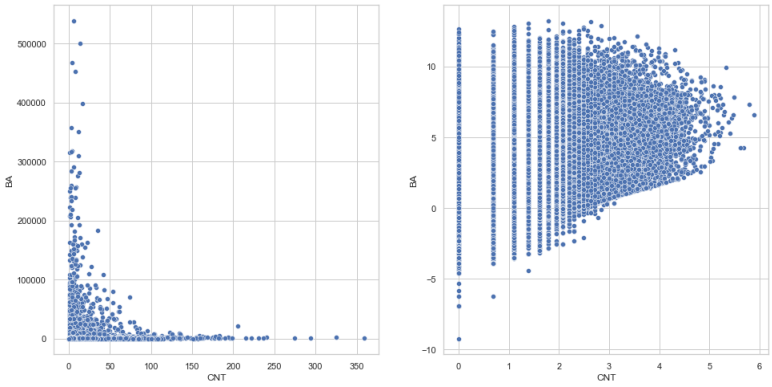

Figure 2: CNT vs BA

We can also see the distribution of the number of fires by year and month [Figure 3]. There is no particular trend for the evolution of the mean number of wildfires per year even if one can think

that global warming may have an effect for example. It could be explained by the improvement of wildfires prediction. We also can see that the months May, June and September have the lowest mean number of wildfires and April and July the highest. This seems a little odd but may be due to the dataset.

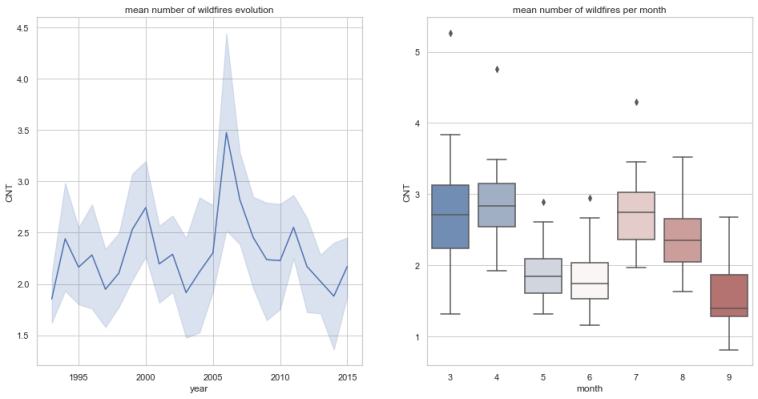

Figure 3: Temporal distribution of CNT

The following figures [Figure 4] represent the histograms of the two response variables we want to predict using only the strictly positive values on the logarithm scale.
The first plot suggests that a Poisson regression can be a good start for the search of the final model. We can use for example a generalized linear model (GLM) with the Poisson distribution and the logarithm as a link function.
The second is very interesting : It looks like a nearly perfect Gaussian distribution on the right side (BA $\geq$ 10) and a very noisy one on the left. This suggests that we could use a log normal distribution for positive values of BA and combine that with a zero/non-zero BA classifier.

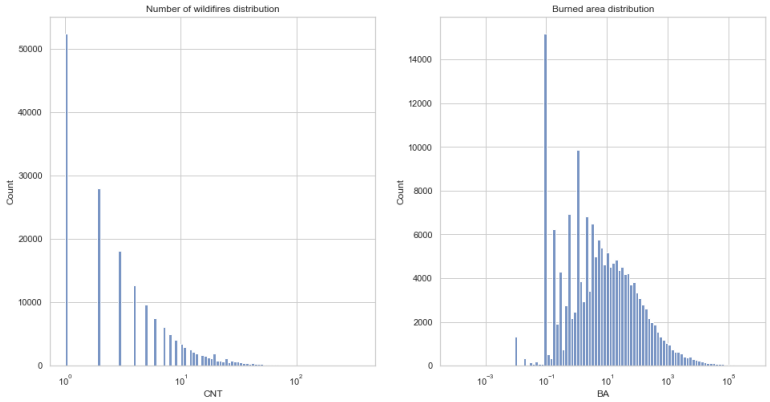

Figure 4: Non-Zero CNT and BA histograms 1

We can have a closer look at the burned area distribution by separating the data into two parts using the threshold BA = 10 [Figure 5]. This suggests using three models : First a classifier for BA with classes BA=0, BA $\in$]0,10] and BA $>$ 10. Then a log normal distribution for BA $\geq$ 10 and some kind of Poisson or Gamma regression for BA $\leq$ 10. All of this thoughts are useful for the next

section where we present the models.

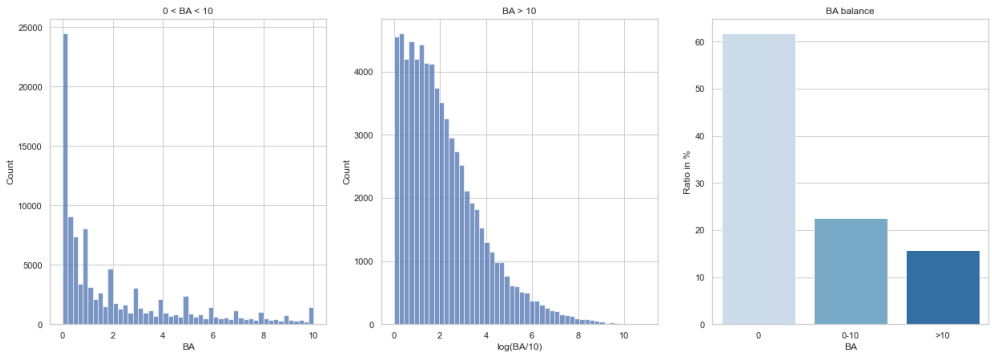

Figure 5: BA histograms

Going back to the actual challenge, there are 28 severity thresholds that represent the levels of the distribution we want to predict. We can look at the histograms [Figure 6] using these thresholds as bins but only using non zero values. It looks looks like the data is highly unbalanced for CNT and BA but at the same time the scoring function of the challenge gives importance to extreme values so we should find a way to improve the modeling of rare events : We could use extreme distributions and we also could use under sampling or over sampling for balancing the data if we plan to use some sort of classification.

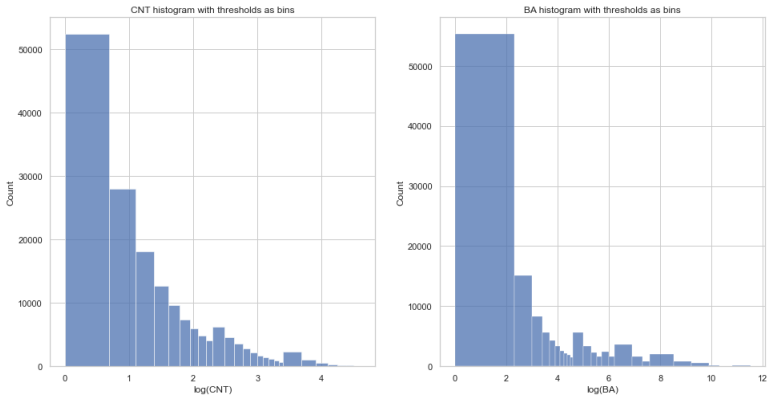

Figure 6: Non-Zero CNT and BA histograms 2

In the last parts, we removed the zeros for the analysis many times without talking about the balance of dataset. It is highly unbalanced since fires are quite "rare" if we look at every part of US (even urban areas with no forest). We actually have 62% of rows with no fires, 73% with one fire or less and 94% with 10 fires or less. As a remark, we previously suggested a Poisson regression for CNT but with such high number of zeros a Zero-Inflated Poisson regression may be better suited. To conclude, wildfires prediction isn't an easy task because of the unbalance of fire occurrences and the huge difference in fire sizes even if extreme values are quite rare. For the next parts, we will have a closer look at the available predictors which are land covers and meteorological variables for each grid cell at a given time.

2.3 Land Covers

In this part, we analyse the land cover variables which are labeled lc1 to lc18 and represent nearly the fraction of a particular type of land cover in the grid cell so their sum for a specific time and place is approximately 1. The description of each variable is as follows :

1. cropland rain fed
2. cropland rain fed herbaceous cover
3. mosaic cropland
4. mosaic natural vegetation
5. tree broad leaved evergreen closed to open
6. tree broad leaved deciduous closed to open
7. tree needle leaved evergreen closed to open
8. tree needle leaved deciduous closed to open
9. tree mixed
10. mosaic tree and shrub
11. shrub land
12. grassland
13. sparse vegetation
14. tree cover flooded fresh or brackish water
15. shrub or herbaceous cover flooded
16. urban
17. bare areas
18. water

We can start by looking at the Pearson's correlation between the land variables [Figure 7]. It doesn't seem that we have a lot of high correlations but it we may have a non linear ones that aren't captured by the Pearson's approach. We feel like we shouldn't use that much variables when fitting models. One possible dimension reduction method is Principal Component Analysis but if we have a more "physics" approach to the problem, we could create a new feature that represents the percentage or fraction of "forest" which is the types of land cover that tends to burn during fires such as shrubs. In our case, it could be just the sum of the appropriate land covers and we should have a value in [0,1].

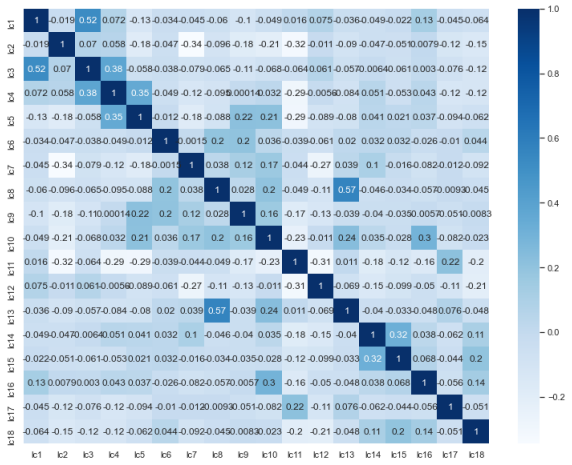

Figure 7: Pearson's correlation for land covers

We also plot the distribution of each land cover against CNT and BA [Figure 8-9]. First, we see that the scales of the variables are not the same even if they are all in [0,1]. Some land covers have a rather small fraction of the land but we believe it doesn't mean they are not important. Also, we can see that some variables have no particular pattern at all so they may be not useful to use in the models or need to be transformed. Some land covers such as lc1, lc3, lc4, lc6, lc13 and lc17 have similar patterns : An increase of the land cover decreases the number of wildfires and the aggregated burned area.

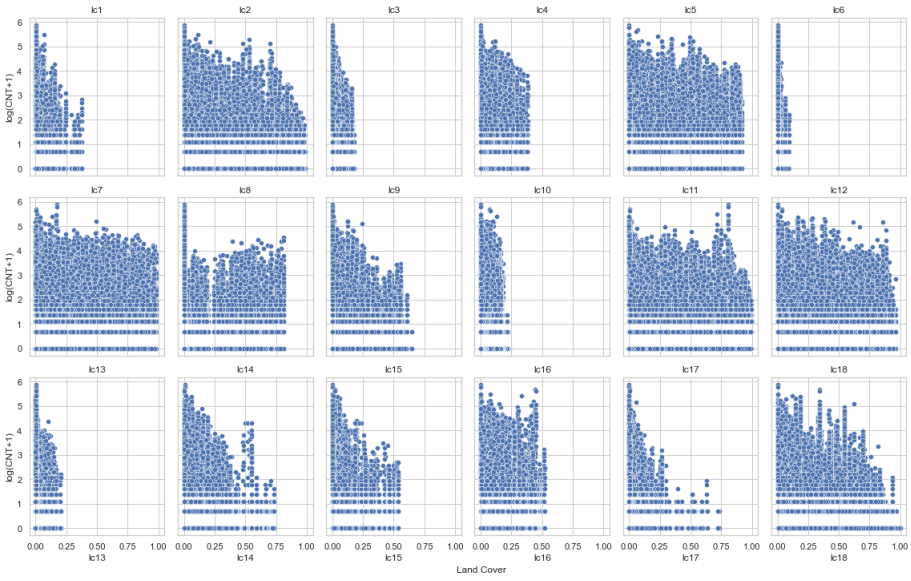

Figure 8: Land covers - CNT

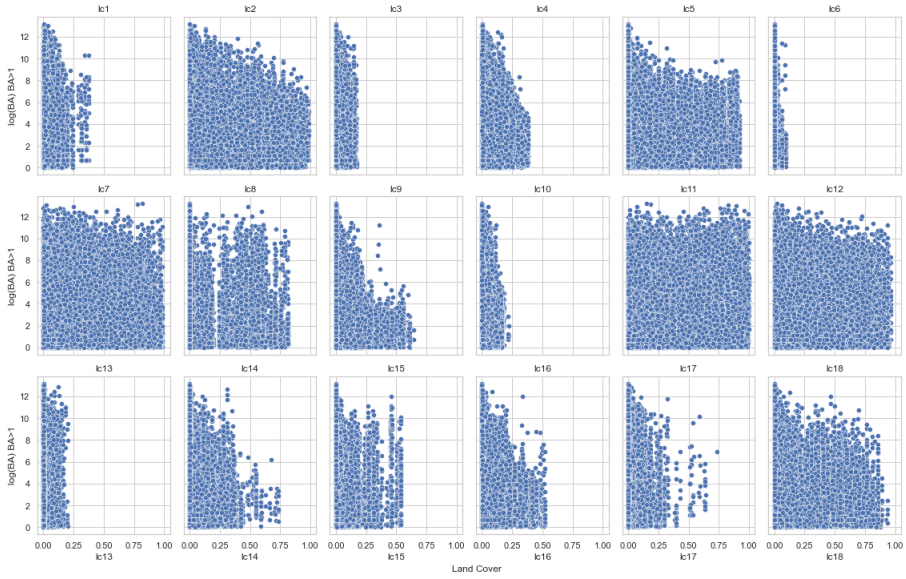

Figure 9: Land covers - BA

Another idea we got is to see what is really happening in land covers during fires : If a certain type of land is a fuel then it should be present before a fire happens and decrease until the fire is ended. So we chose the grid cell with the highest average burned area (longitude -115.25 and latitude 42.25) and plotted the time series of BA and each land cover and got the following figure [Figure 10]:

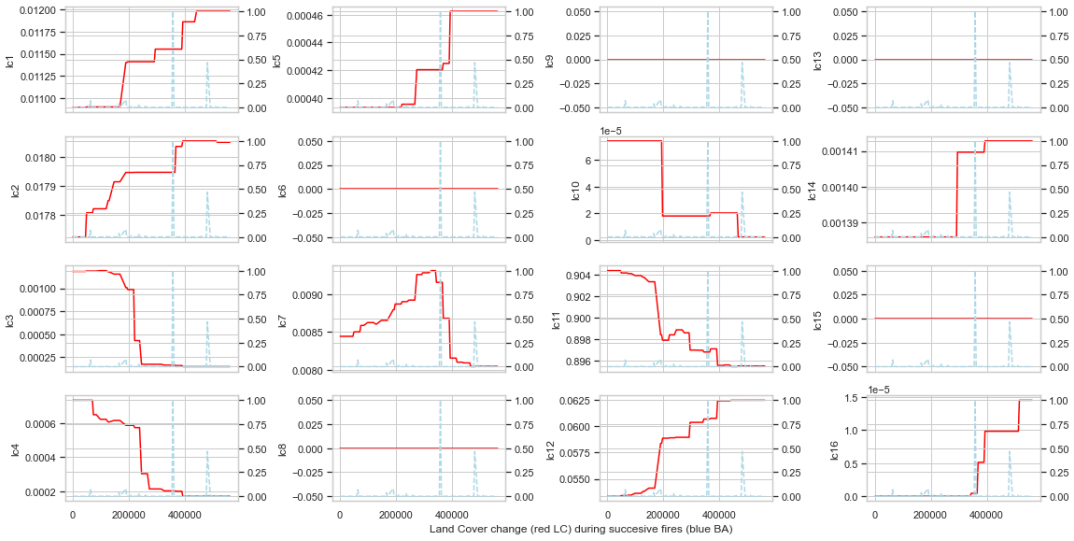

Figure 10: Land covers and burned area evolution

First, lc6, lc8, lc9, lc13 and lc15 are not present and don't change in time so they might have no effect. The interesting land covers are lc7 and lc11 : The land cover lc11 represents shrub land which is a mature vegetation type in a particular region and remain stable over time, or a transitional

community that occurs temporarily as the result of a disturbance, such as fire. A stable state may be maintained by regular natural disturbance such as fire or browsing. Shrub land is known to be unsuitable for human habitation because of the danger of fire.
Speaking of population, we plotted the spatial distribution of urban density (lc16) across the US and it seems there is a correlation between urban density and wildfires since most of the fires start by a human error or because of ecological impacts on the area from construction. (I SAW A PAPER SAY 90% OF FIRES ARE DUE TO HUMAN ACTIVITY BUT CAN T FIND IT. add the reference if you can)

2.4 Meteorological Variables

In this part, we analyse the meteorological variables which are labeled clim1 to clim10. The description of each variable is as follows :

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/m2)
7. Surface net thermal radiation (net flux of longwave radiation; mostly radiation emitted by the surface) (J/m2)
8. Surface pressure (Pa)
9. Evaporation (of water) (m)
10. Precipitation (m)

We can start by looking at the Pearson's correlation between the climate variables [Figure 11]. There are a lot of highly correlated meteorological variables, some of them are expected based on physics knowledge such as the two altitude related variables (0.7), the two temperature variables (0.76) and the mean altitude and surface pressure (-1). This needs some serious investigation and a dimension reduction before fitting models.

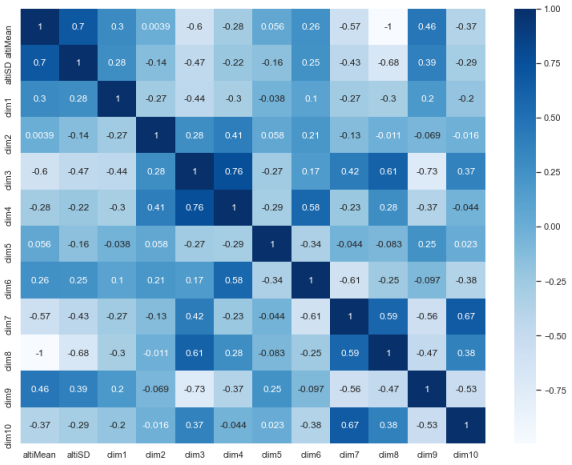

Figure 11: Pearson's correlation for climate variables

We also plot the distribution of each climate variable against CNT [Figure 12]. We use the min-max scaler for each variable to have the same scale order ([0,1]).

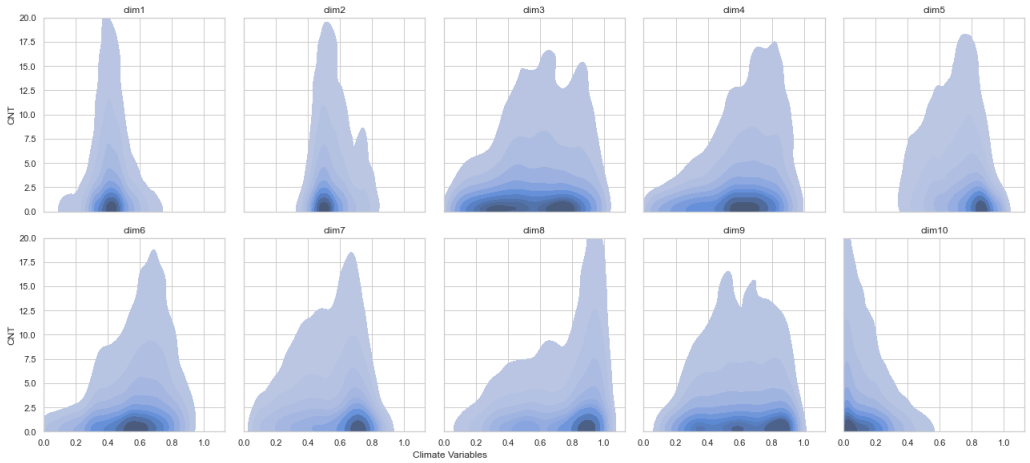

Figure 12: Meteorological Variables - CNT

We can see that precipitation (clim10) has a high density around zero CNT which is reasonable. The variables clim4, clim7 and clim8 share the same pattern where we have a higher number of fires for high values of these variables. The wind related variable clim1 and clim2 are nearly the same which means that the direction of wind doesn't matter but only the value does. One observation is the variables that seem correlated with CNT are used in the Fire Weather Index (FWI) [1], However, they are usually not used in a linear form. We could get inspired by the FWI to transform the variables.

2.5 New Variables and Scaling

As we explained before, the direction of the wind doesn't matter so we replaced the variables clim1 and clim2 by a new feature "wind" which is the root mean square of the two variables. We

also added the variable "RH" which is the relative humidity computed as an approximation using the available data [2]. We added a third variable "meanCNT" which represents the mean number of fires for each grid cell at a specific month which is computed on the training set and copied to the validation set since we believe some regions have more common fires based on unavailable factors which could be captured by the mean.
The land covers variables are in [0,1] so we can keep them in that scale. For the climate variables, the scales are quite different so we decided to standardize them such that they have mean 0 and standard deviation 1.

2.6 Validation Set

No data have been masked for uneven years (1993, 1995, ..., 2015) but only for even years (1994, 1996, ..., 2014). First, as we said before there are 80000 observations of CNT and BA respectively we want to predict. For each each variable, we have 39% of the second one available to use. Second, we don't have any particular pattern in terms of variables for the validation [Figure 13] : It looks that it uniformly distributed despite that the organizer of the challenge said "The spatial and temporal positions of validation data are not completely random, but they tend to be clustered in space and time" [3]. We may have missed an important aspect so our work could be improved later.

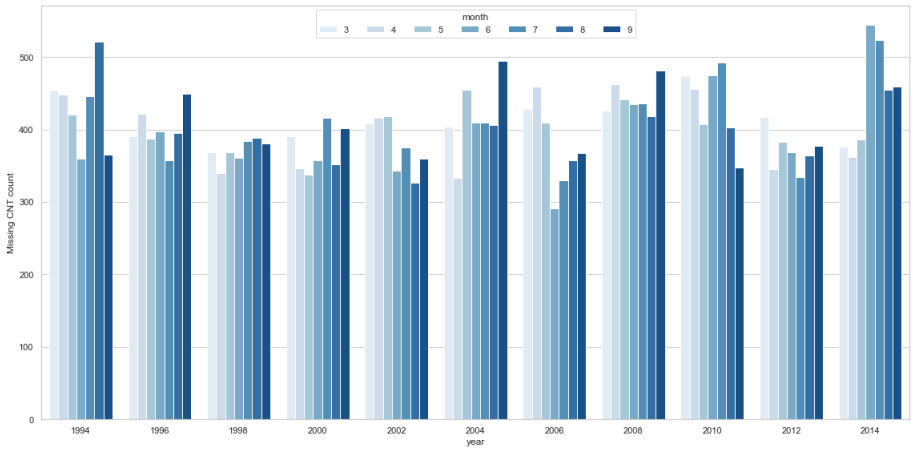

Figure 13: Temporal distribution of the validation set

We finally decided to use the years 1995, 2005 and 2015 for the training set and 1997 and 2007 for the testing set. This makes the models quicker to fit in order to compare several models and implementations. However, we usually use all the available data to fit the final model for the challenge. This is not a rule for every tested model but at each step in the following parts of this report we should clearly explain which sample of the data we use to fit and test the model.

References

[1] NWCG. *Fire Weather Index (FWI)*. URL: https://www.nwcg.gov/publications/pms437/cffdrs/fire-weather-index-system.

[2] Haiganoush K.Preisler. "Probability based models for estimation of wildfire risk". In: (2004).

[3] Thomas Opitz. *Wildfires and their extremes: a global challenge*. 2021. URL: https://www.maths.ed.ac.uk/school-of-mathematics/eva-2021/competitions/data-challenge.