INSA ROUEN NORMANDIE

PROJET DE SCIENCE DES DONNÉES

Le déclenchement d'un feu de forêt est-il prévisible ?

Adrien Pasquesoone
Nathan Lesourd

Monsieur Benoit Gaüzère

STPI 2 - Année 2020/2021

Table des matières

Section 1

Introduction

Au cours de ces dernières années, le nombre de feux de forêt n'a cessé d'augmenter à travers le monde. Cette augmentation a atteint son paroxysme en 2019 avec des incendies historiques en Australie et en Sibérie. Pas un mois ne s'est passé en 2019 sans que des incendies hors normes ne ravagent des milliers d'hectares de forêt. Juin 2019, l'Australie est victime du plus gros feu de forêt de son histoire. Cinq millions d'hectares seront détruits au cours des 8 mois d'incendie. Aout 2019, l'Amazonie connait l'un des plus importants feux de forêt qu'elle n'ait jamais connu. Dans le même temps, 3 millions d'hectares de forêts russes partent en fumée. On pourrait aussi parler des incendies sur les collines californiennes ou en Indonésie pendant l'automne 2019 mais le constat est déjà assez lourd pour constater l'ampleur des feux de forêts.

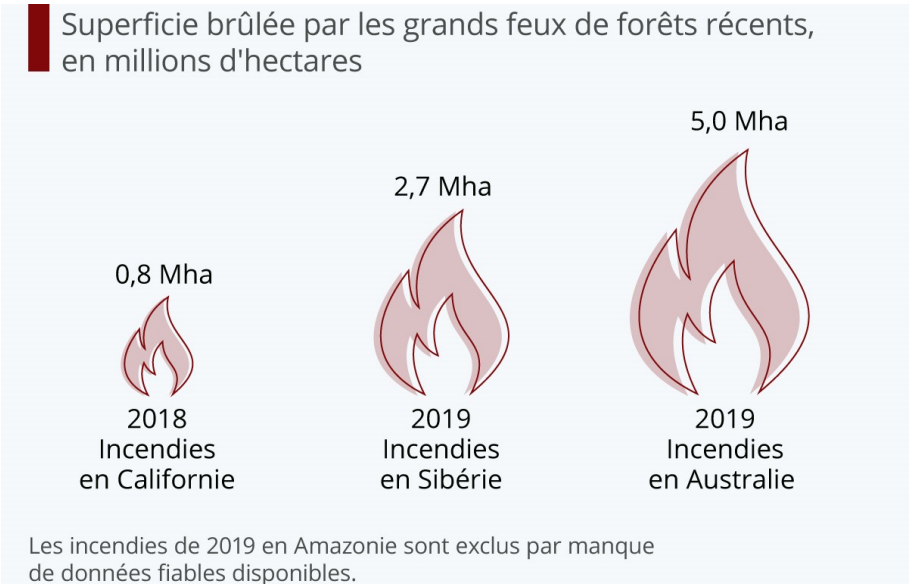

FIGURE 1 – La taille démesurée des incendies en Australie - statista

Il est d'autant plus important de prendre en considération les feux de forêt qui ont un impact très négatif sur notre écosystème. Lors de ces incendies la faune et la flore sont très gravement affectées avec la destruction de la végétation et la disparition d'espèces. On peut citer le Koala d'Australie et l'orang-outan d'Indonésie qui sont en voie d'extinction. De plus, la combustion du bois rejette énormément de CO2 ce qui accélère le réchauffement climatique. Selon le Global Carbon Atlas les feux de forêts ont produit l'émission de 6375 millions de tonnes de CO2 en 2019. C'est plus que l'émission annuelle des Etats-Unis avec 5312 millions de tonnes de CO2.

Tous ces exemples et chiffres ont permis de mettre en avant la gravité de la situation. Mais comment peut-on réduire le nombre de feux de forêts ? Dépendent-ils de facteurs ou sont-ils juste aléatoires ? Peut-on les anticiper ? C'est tout l'objet de cette étude.

Section 2

Description des données

2.1 Présentation des données

Notre base de données est constituée d'informations de feux de forêt dans un parc au Nord du Portugal. Ces informations sont regroupées en 13 variables pour 517 observations. Pour mieux se représenter ces données nous avons seulement représenté ci dessous les 20 premières observations.

X	Y	Month	Day	FFMC	DMC	DC	ISI	Temp	RH	Wind	Rain	Area
7	5	3	5	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
7	4	10	2	90.6	35.4	669.1	6.7	18	33	0.9	0	0
7	4	10	6	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	3	5	91.7	33.3	77.5	9	8.3	97	4	0.2	0
8	6	3	7	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
8	6	8	7	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	6	8	1	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
8	6	8	1	91.5	145.4	608.2	10.7	8	86	2.2	0	0
8	6	7	2	91	129.5	692.6	7	13.1	63	5.4	0	0
7	5	7	6	92.5	88	698.6	7.1	22.8	40	4	0	0
7	5	7	6	92.5	88	698.6	7.1	17.8	51	7.2	0	0
7	5	7	6	92.8	73.2	713	22.6	19.3	38	4	0	0
6	5	8	5	63.5	70.8	665.3	0.8	17	72	6.7	0	0
6	5	9	1	90.9	126.5	686.5	7	21.3	42	2.2	0	0
6	5	9	3	92.9	133.3	699.6	9.2	26.4	21	4.5	0	0
6	5	9	5	93.3	141.2	713.9	13.9	22.9	44	5.4	0	0
5	5	3	6	91.7	35.8	80.8	7.8	15.1	27	5.4	0	0
8	5	10	1	84.9	32.8	664.2	3	16.7	47	4.9	0	0
6	4	3	3	89.2	27.9	70.8	6.3	15.9	35	4	0	0
6	4	4	6	86.3	27.4	97.1	5.1	9.3	44	4.5	0	0

FIGURE 2 – Tableau des données (20 observations)

Nous voyons donc les observations décrites par 13 variables :

- ***X :*** La coordonée selon l'axe x
- ***Y :*** La coordonée selon l'axe y

Ci-dessous, voici la carte du parc qui permet de mieux s'imaginer ces variables géographiques

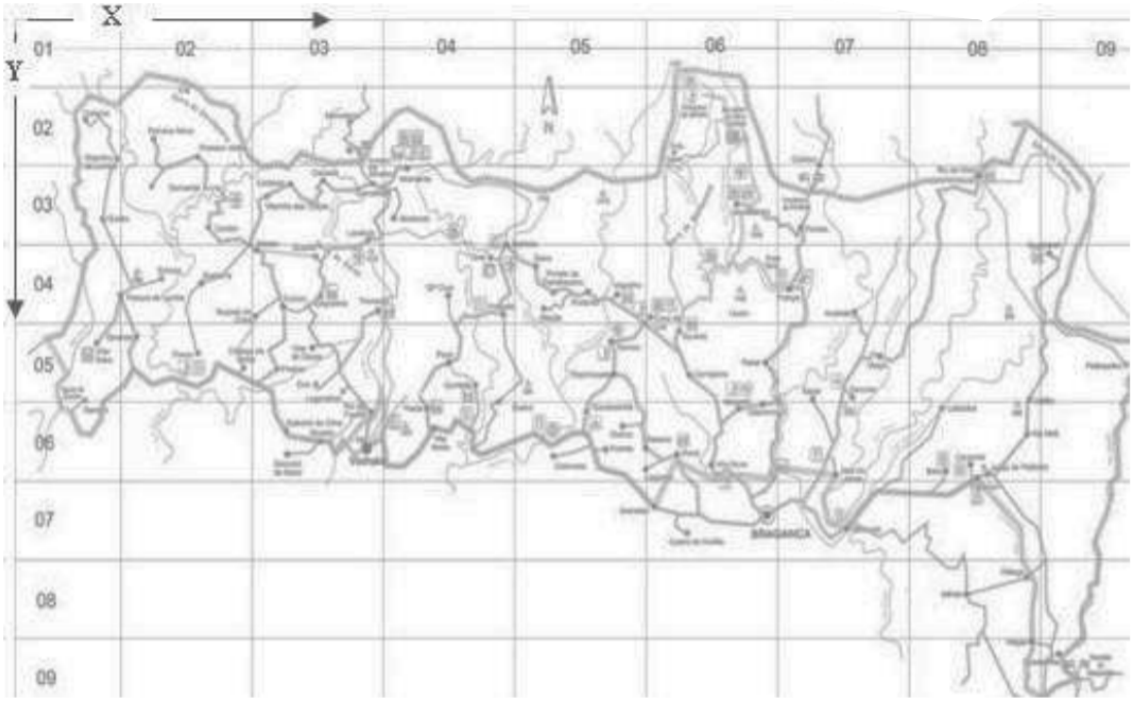

FIGURE 3 – Carte du parc naturel de Montesinho

Il y a en plus des variables géographiques, des variables temporelles :

- ***Le mois***
- ***La date***

Ensuite il y a les informations relatives aux taux d'humidité :

- ***FFMC :*** Le code d'humidité du combustible fin.
- ***DMC :*** Le DMC représente l'humidité du carburant des matières organiques décomposées sous la lisière.
- ***DC :*** Le code de sécheresse (DC) représente le séchage en profondeur dans le sol.
- ***ISI :*** Le ISI intègre l'humidité du combustible pour les combustibles morts fins et la vitesse du vent en surface pour estimer un potentiel de propagation.

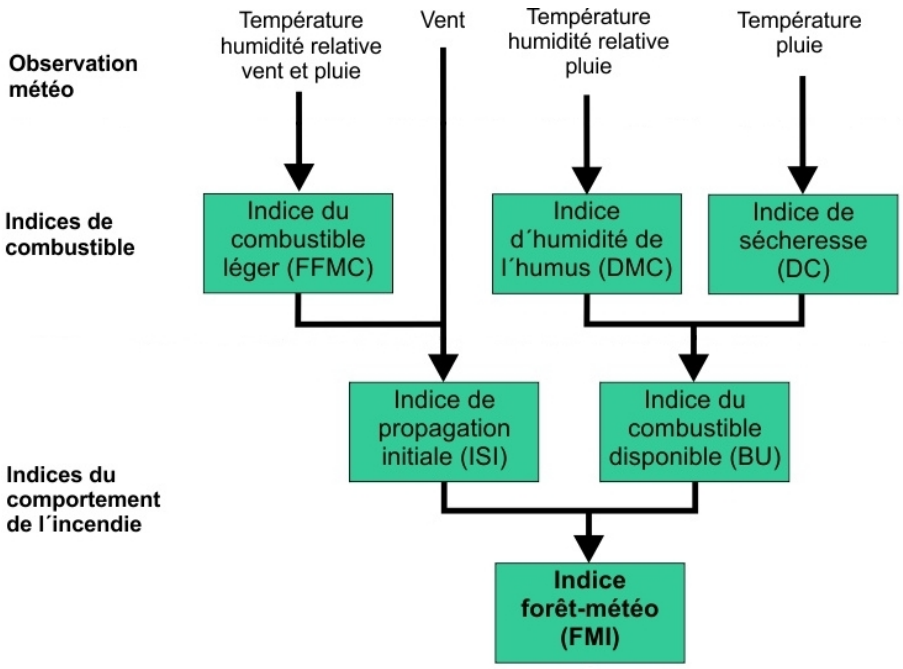

FIGURE 4 – Schéma de principe de l'indice forêt-météo

Les prochaines variables sont les données météorologiques :

- ***La température (en °C)***
- ***RH (en %) :*** Le taux d'humidité.
- ***Le vent (en km/h) :*** La vitesse du vent.
- ***La pluie (en mm/m²) :*** La pluviométrie.

Enfin la dernière variable représente la surface brûlée dans le parc lors du feu :

- ***La surface :*** en hectare.

2.2 Boîtes à moustaches

Continuons la description de nos données en réalisant les boîtes à moustaches des 13 différentes variables. Cela va nous permettre d'avoir une vue d'ensemble des valeurs que peuvent prendre les variables et de déceler une quelconque aberration.

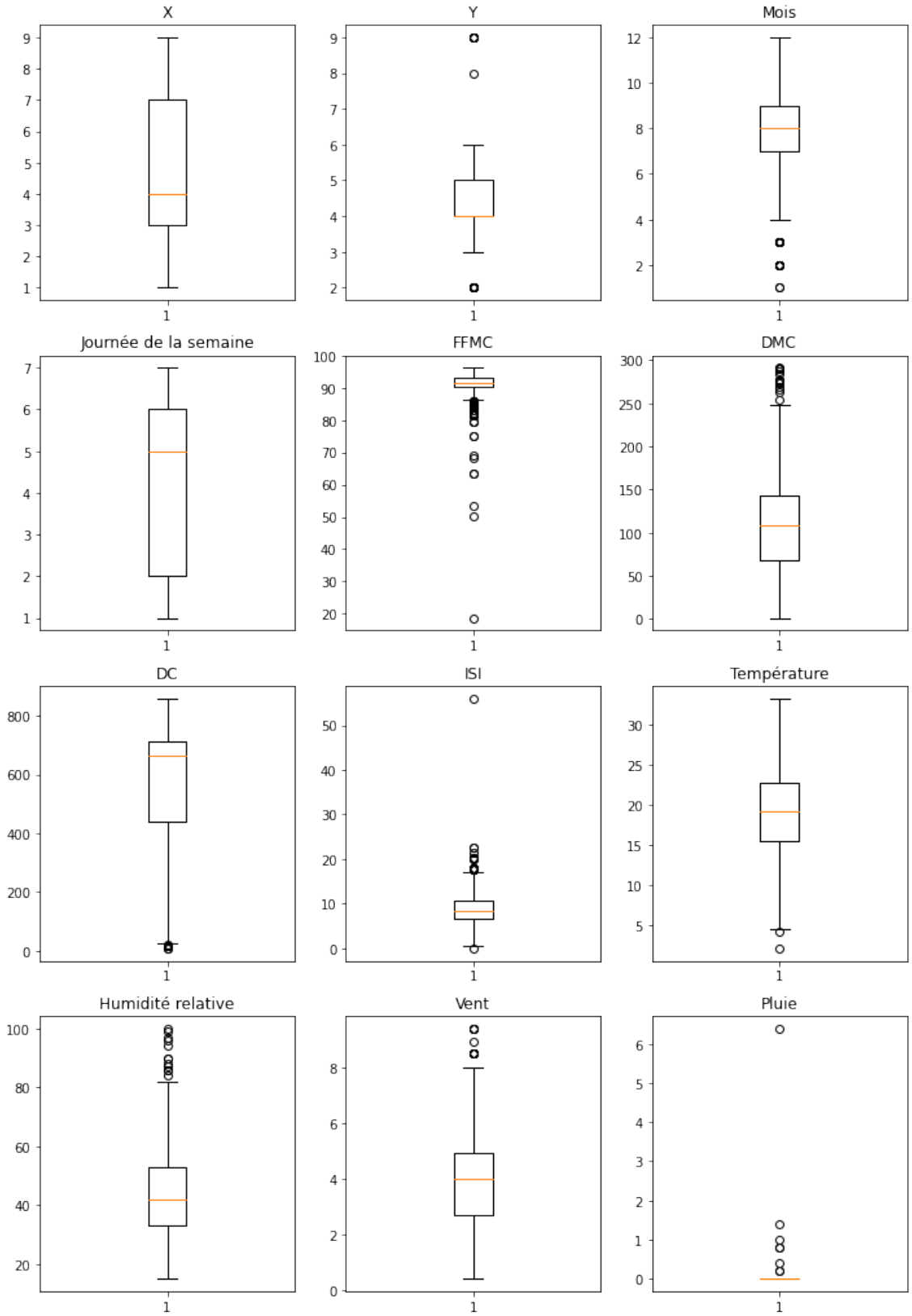

FIGURE 5 – Boites à moustaches des 12 premières variables

Tout d'abord intéressons nous aux deux premières variables géographiques. On constate que les feux de forêts se déclenchent de façon homogène horizontalement (de l'Ouest vers l'Est du parc) mais de façon plus concentrée verticalement, surtout dans le sud. Ainsi la latitude pourrait impacter le déclenchement d'un feu.

Ensuite nous retrouvons deux variables temporelles avec le mois et la journée de la semaine. 75% des incendies se sont déclenchées entre juillet et septembre correspondant à l'été au Portugal. Cette saison est propice puisque les températures sont plus élévées et la végétation est sèche. On ne peut rien conclure quant à la journée de la semaine la plus propice aux incendies.

Passons aux variables relatives à l'humidité. On remarque que les valeurs de l'indice FFMC se rapproche très fortement de la valeur moyennne, égale à 90. Selon la figure 19 présente en annexe, 90 représente un indice très haut. L'humidité du combustible léger étant faible, le feu peut prendre rapidement. Toujours selon la figure 19, le DMC et le DC sortent du lot par leur valeur moyenne considérée comme extrême. Ces derniers indices sont pertinents puisque leur valeur extrême est synonyme d'un incendie imminent.

Enfin, les variables météorologiques nous donnent plusieurs informations. L'incendie se déroule plutot lorsqu'il fait chaud avec une température moyenne de 20°C. On aurait pu s'attendre à une température moyenne plus élévée du fait que les incendies se déroulent dans un parc au Portugal et le plus souvent en été (la température moyenne à Lisbonne en juillet est de 30°C). Le taux d'humidité moyen est plutot élévé tandis que la vitesse du vent moyen et la pluviométrie sont plutôt faibles.

D'après ces premières données, l'incendie a lieu le plus souvent en été dans la partie sud du parc. Les indices relatifs à l'humidité sont tout à fait pertinents pour notre étude. La pluviométrie doit-être faible voire nulle pour que l'incendie puisse partir.

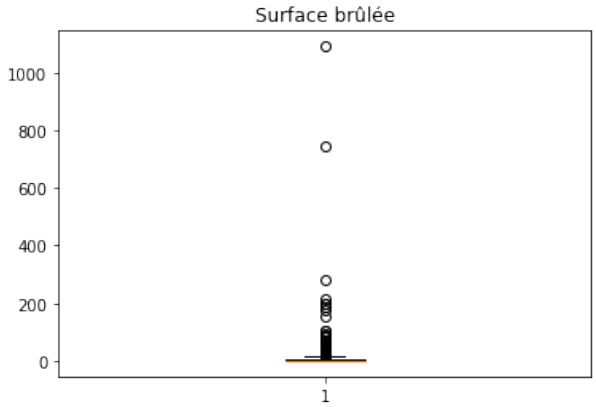

Figure 6 – Boites à moustaches de la surface brûlée

Voici maintenant la boîte à moustaches de l'une des variables les plus importantes de notre projet. En effet, on s'est interessé précédemment aux différents paramètres de la situation lors du feu. Mais la surface brulée quant à elle quantifie l'ampleur et les dégats du feu. On constate deux gros feux de forêt avec pas moins de 1100 hectares pour l'un et 750 hectares pour l'autre. La surface moyenne brulée est bien plus faible, 13 hectares. La plupart des feux sont donc assez petits et il serait interesant de regarder les conditions météorologiques lors des deux immenses feux.

2.3 Analyse en Composantes Principales (ACP)

Nous avons ensuite fait l'ACP afin de nous rendre compte des variables importantes dans la descrpition des feux de forêt. Avec le graphe ci-dessous nous pouvons remarquer que seulement 3 variables sur 13 suffisent pour représenter un peu moins de 50% des informations.

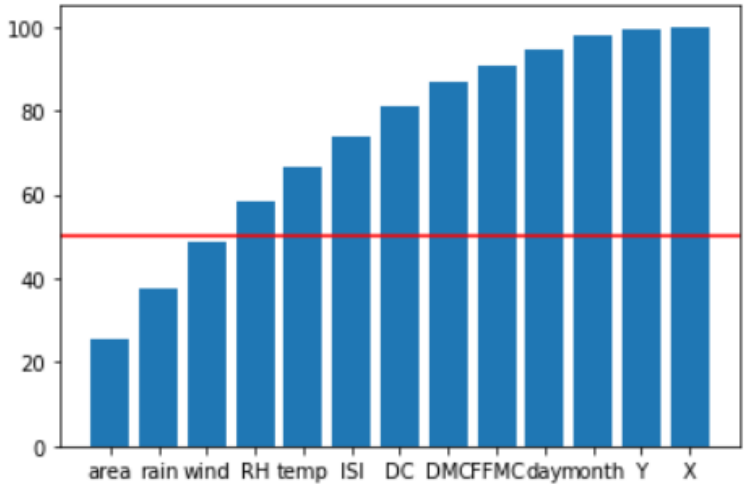

Figure 7 – Importance des variables

Nous avons aussi fait la projection et la reconstruction des données mais cela ne nous semblait pas pertinent de le présenter (voir annexe). Ensuite grâce à l'ACP et à l'analyse des valeurs propres nous constatons que les variables qui contiennent le plus d'informations par rapport aux feux de forêts sont la surface brûlée, le taux de pluie et enfin la vitesse du vent.

	X	Y	month	day	FFMC	DMC
Valeurs propres	0.0791783	0.212075	0.435427	0.461088	0.523217	0.756939

DC	ISI	temp	RH	wind	rain	area
0.926435	0.984627	1.06437	1.23443	1.43763	1.56939	3.31519

Figure 8 – Valeurs Propres

Section 3

Régression

Voici ci-dessous la matrice de corrélation, nous pouvons observer que la plupart des variables ne sont pas liées entre elles deux à deux. Prenons par exemple la surface brûlée avec la température ayant un coefficient de corrélation de 0.098.

	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
X	1	0.54	-0.071	-0.028	-0.021	-0.048	-0.086	0.006	-0.051	0.085	0.019	0.065	0.063
Y	0.54	1	-0.07	-0.007	-0.046	0.008	-0.101	-0.024	-0.024	0.062	-0.02	0.033	0.045
month	-0.071	-0.07	1	-0.054	0.289	0.469	0.866	0.183	0.37	-0.096	-0.091	0.014	0.058
day	-0.028	-0.007	-0.054	1	-0.038	0.065	-0	0.038	0.051	0.093	0.032	-0.049	0.023
FFMC	-0.021	-0.046	0.289	-0.038	1	0.383	0.331	0.532	0.432	-0.301	-0.028	0.057	0.04
DMC	-0.048	0.008	0.469	0.065	0.383	1	0.682	0.305	0.47	0.074	-0.105	0.075	0.073
DC	-0.086	-0.101	0.866	-0	0.331	0.682	1	0.229	0.496	-0.039	-0.203	0.036	0.049
ISI	0.006	-0.024	0.183	0.038	0.532	0.305	0.229	1	0.394	-0.133	0.107	0.068	0.008
temp	-0.051	-0.024	0.37	0.051	0.432	0.47	0.496	0.394	1	-0.527	-0.227	0.069	0.098
RH	0.085	0.062	-0.096	0.093	-0.301	0.074	-0.039	-0.133	-0.527	1	0.069	0.1	-0.076
wind	0.019	-0.02	-0.091	0.032	-0.028	-0.105	-0.203	0.107	-0.227	0.069	1	0.061	0.012
rain	0.065	0.033	0.014	-0.049	0.057	0.075	0.036	0.068	0.069	0.1	0.061	1	-0.007
area	0.063	0.045	0.058	0.023	0.04	0.073	0.049	0.008	0.098	-0.076	0.012	-0.007	1

FIGURE 9 – Matrice de corrélation

C'est pour cela que la régression suivante n'est pas satisfaisante. Le modèle que nous obtenons ne nous permet pas de conclure d'hypothèses.

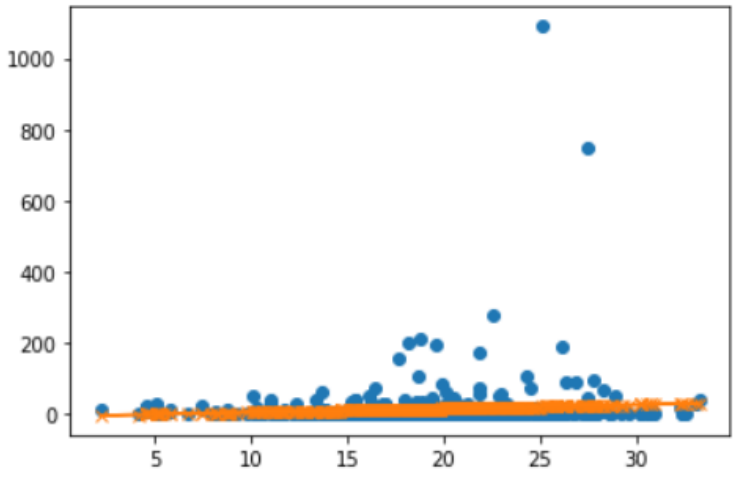

FIGURE 10 – Régression de la surface brûlée en fonction de la température

Néanmoins, on peut répertorier 4 coefficients de corrélation supérieur à 0.5 dans le tableau suivant :

Variable 1	Variable 2	Cofficient corrélation
DC	DMC	0,682
Y	X	0,54
ISI	FFMC	0,532
RH	Temp	-0,527

FIGURE 11 – Plus importants coefficients de corrélation

Regardons si ces régressions simples incluant deux variables sont correctes et si l'on peut en tirer des conclusions. Pour la suite de cette étude, on cherche à exprimer la variable 1 en fonction de la variable 2.

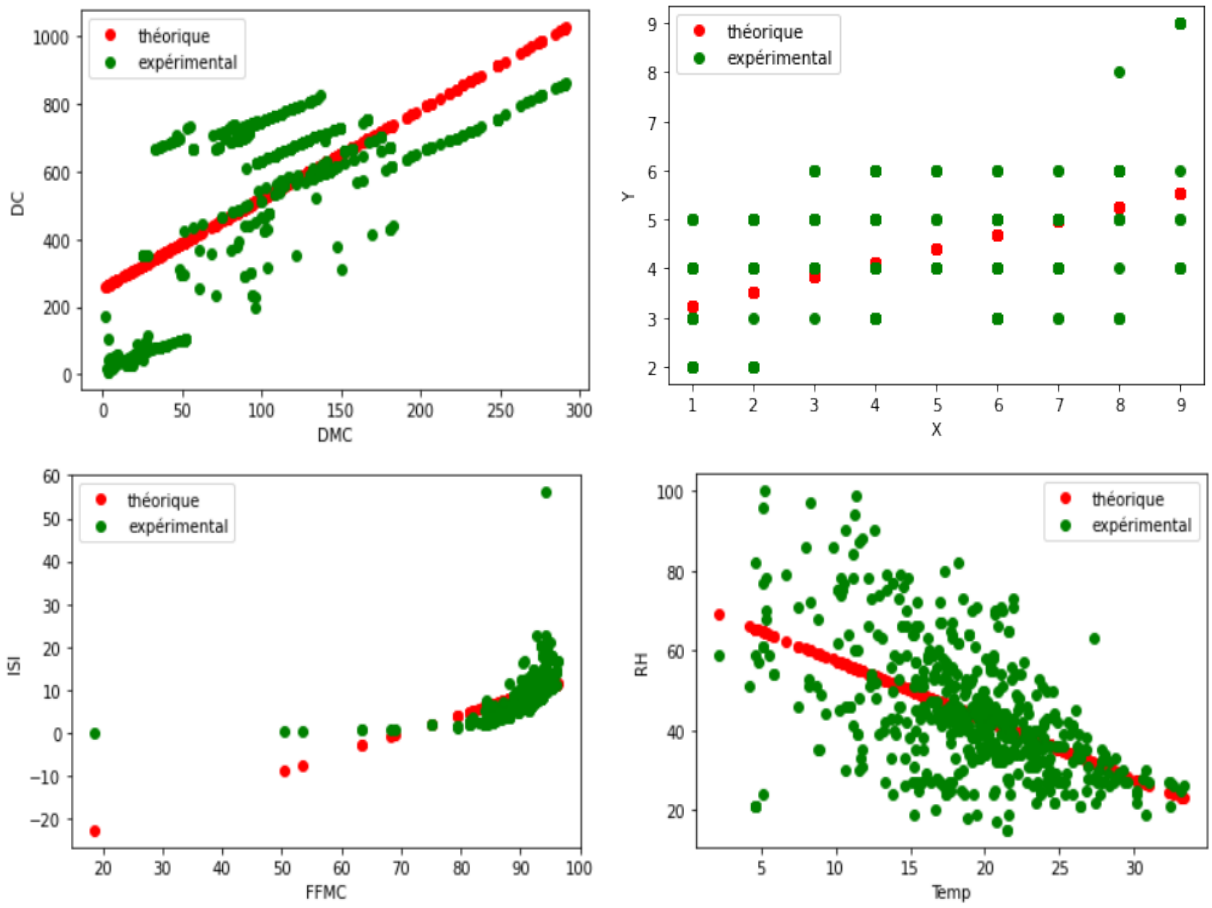

FIGURE 12 – Graphiques des différentes régressions

On remarque que les comportements de ces régressions sont tout à fait différents. Si l'on considère le premier graphique avec le coefficient de détermination le plus important, on voit que notre regression est colinéaire avec différentes droites du plan. Il n'existe pas une relation linéaire entre DC et DMC mais bien plusieurs. On pouvait s'y attendre avec la figure 4 puisque

ces deux indices s'appuit sur des facteurs communs (température et pluie). Ensuite on s'est interessé à la relation entre l'abcisse et l'ordonnée d'un feu. Une relation linéaire se dégage mais sans grande précision, dû au large nuage autour de la droite de régression. En effet, la description de la trajectoire semble difficile à prévoir tant elle dépend de paramètres incertains. La troisième régression est interessante de par la forme du nuage de points des différents feux. On reconnait le début d'une fonction exponentielle. Enfin la dernière régression est convaincante, l'humidité est inversement proportionnelle à la température.

a	b	Covariance	Coef R2	Résidus	H	Contribution
2.64228	254.984	32962.4	0.465385	3.30725e-13	0.00386847	0.00209372
0.2868	2.96067	1.07438	0.291112	-1.44308e-16	0.00386847	0.00234002
2.1841e-41	8.22392	16.8598	0.190571	-3.76574e-15	0.00386847	0.00264255
-1.48204	72.2828	192.575	0.278141	9.67548e-15	0.00386847	0.00242215

Figure 13 – Tableau de données des régressions

Le précédent tableau synthétise les différentes grandeurs de notre régression. La régression possédant le meilleur coefficient de détermination est la première. Néanmoins, rien de glorieux puisqu'elle n'exède même pas les 0,5. Ainsi, on ne peut pas conclure de relation linéaire mais uniquement des corrélation qui suivent une droite affine.

Comme dit ci-dessus, nous avons constaté que la régression ISI-FMC s'apparentait à un modèle exponentielle. Nous avons donc appliqué ce modèle et voici le résultat.

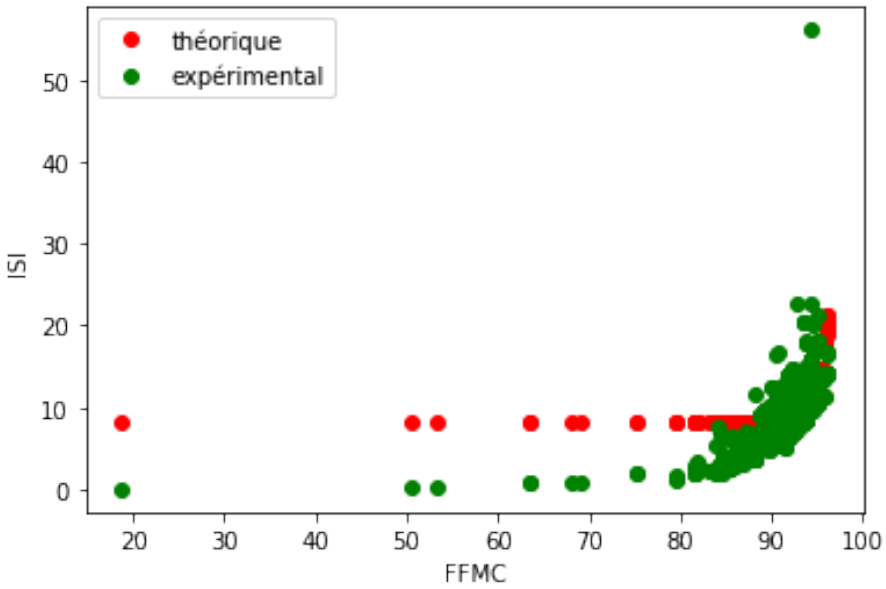

Figure 14 – Régression ISI-FMC modèle exponentielle

Cependant, nous obtenons un coefficient R2 égal à 0.19, ce qui est très peu satisfaisant. Nous pensons que l'approche exponentielle est une bonne approche mais que la méthode que nous avons utilisé (changement de variable) n'est pas appropriée.

Modèle utilisé :

$$\alpha * exp(x) + \beta$$

Modèle peut-être plus approprié :

$$\alpha * exp(x * \gamma) + \beta$$

Enfin, nous avons essayé de faire plusieurs régressions en ayant comme variable à expliquer l'aire de la surface brulée en fonction des douzes autres variables. En effet, cette variable quantifie l'ampleur des dégats du feu, il est donc intéressant d'établir des relations entre la surface et les autres.

	a	b	Covariance	Coef R2	Résidus	H	Contribution
X	1.74383	4.70491	4043.62	0.0040177	3.51835e-15	0.00386847	0.00193111
Y	2.3225	2.861	4051.76	0.00201361	1.53928e-15	0.00386847	0.00201664
Mois	1.62838	0.702138	4046.21	0.00337893	3.07856e-15	0.00386847	0.001295
Journée	0.711785	9.81429	4057.76	0.000533991	2.85866e-15	0.00386847	0.0015881
FFMC	0.462672	-29.0914	4053.4	0.00160978	3.08956e-14	0.00386847	0.00127624
DMC	0.0725491	4.80361	4038.3	0.00532817	-8.79589e-16	0.00386847	0.00188318
DC	0.0126721	5.90372	4050.03	0.0024387	4.83774e-15	0.00386847	0.00127549
ISI	0.115287	11.8072	4059.65	6.81894e-05	2.19897e-15	0.00386847	0.00134718
Température	1.07263	-7.41375	4021.06	0.00957347	-4.83774e-15	0.00386847	0.00224971
Humidité	-0.294604	25.8948	4036.78	0.00570305	2.41887e-15	0.00386847	0.00194718
Vent	0.437622	11.0891	4059.32	0.000151715	3.07856e-15	0.00386847	0.00116381
Pluie	-1.58424	12.8816	4059.71	5.4254e-05	2.19897e-15	0.00386847	0.00243643

FIGURE 15 – Tableau de données des douze régressions simples

On constate que l'ensemble des coefficients de corrélation sont décevants avec une covariance énorme (en moyenne 4000). Ces résultats amènent à rejeter les douzes modèles linéaires.

Section 4

Tests Statistiques

Dans cette dernière partie, nous allons aborder les tests statistiques. Nous avons décidé de faire plusieurs test de Student afin de comparer plusieurs variables quantitatives :

- ***La surface de forêt brûlée***
- ***La température, l'humidité, le mois***

Pour effectuer ce test on émet 2 hypothèses :

- ***H0 :*** la température et la surface brûlée sont indépendantes.
- ***H1 :*** la température et la surface brûlée sont dépendantes.

Voici ci-dessous le code permettant d'effectuer ce test.

```
#H0 : surface brûlée indépendante de la température
#H1 : surface brûlée dépendante de la température
from scipy.stats import ttest_ind
x = data[:,12] # on charge les donnéees de la surface brûlée
y = data[:,8] # on charge les données de la température
ttest_ind(x,y) # test de Student pour données Quantitatif VS Quantitatif
```

```
Ttest_indResult(statistic=-2.1492136924048335, pvalue=0.03184946709710879)
```

FIGURE 16 – Test de Student pour la pluie

On a utilisé la fonction ttest_ind de la bibliothèque scipy.stats afin d'avoir un test de Student bilatéral. On peut remarquer que la p-valeur est égale à 0,0318. Si on prend une marge d'erreur α égal à 5 % on peut déduire que p-valeur < α . On conclue donc que l'hypothèse H0 est rejetée.

On s'intéresse maintenant à la dépendance entre la surface brûlée et l'humidité. Pour effectuer

ce test on émet 2 hypothèses :
https ://fr.overleaf.com/project/60ba3377e26fbd48e68ebbd4***H0 :*** l'humidité et la surface brûlée sont indépendantes. ***H1 :*** l'humidité et la surface brûlée sont dépendantes.

On utilise donc encore le test de Student, vu ci-dessous :

```
#H0 : surface brûlée indépendante de l'humidité
#H1 : surface brûlée dépendante de l'humidité
from scipy.stats import ttest_ind
x = data[:,12] # on charge les donnéees de la surface brûlée
y = data[:,9] # on charge les données de l'humidité'
ttest_ind(x,y) # test de Student pour données Quantitatif VS Quantitatif
```

```
Ttest_indResult(statistic=-10.878845358296164, pvalue=3.596044049353644e-26)
```

FIGURE 17 – Test de Student pour l'humidité

On peut remarquer que la p-valeur est presque égale à 0. Si on prend une marge d'erreur α égal à 5 % on peut déduire que p-valeur < α . On conclue donc que l'hypothèse H0 est rejetée encore une fois.

On s'intéresse maintenant à la dépendance entre la surface brûlée et le mois de l'année. Pour

effectuer ce test on émet 2 hypothèses :

- ***H0 :*** le mois et la surface brûlée sont indépendants.
- ***H1 :*** le mois et la surface brûlée sont dépendantss.

On utilise donc encore le test de Student, vu ci-dessous :

```
#H0 : surface brûlée indépendante du mois de l'année
#H1 : surface brûlée dépendante du mois de l'année
from scipy.stats import ttest_ind
x = data[:,12] # on charge les donnéees de la surface brûlée
y = data[:,2] # on charge les données du mois
ttest_ind(x,y) # test de Student pour données Quantitatif VS Quantitatif
```

```
Ttest_indResult(statistic=1.9236630230674712, pvalue=0.05467165770518388)
```

FIGURE 18 – Test de Student pour le mois

On peut remarquer que la p-valeur est égale à 0,0547. Si on prend une marge d'erreur α égal à 5 % on peut déduire que p-valeur > α . On conclue donc que l'hypothèse H0 acceptée, le mois et la surface de forêt brûlée sont indépendants.

Pour finir, comme on aurait pu s'y attendre la température et l'humidité ont donc une influence sur la surface brûlée. Cependant, le mois n'influe pas sur la surface brûlée. Il ne faut tout de même pas oublier que ce sont des statistiques, ce n'est donc pas forcément la vérité.

Section 5

Conclusion

Pour conclure, à travers ce projet et l'analyse des données sur les feux de forêts nous avons appris certaines informations. Nous arrivons maintenant à mieux cerner ce qui peut déclencher un feu de forêt ou ce qui peut le rendre le plus dévastateur. Nous avons notamment appris que les feux étaient plus propices à se déclencher en été dans le sud du parc du Portugal. Ce qui peut influencer aussi un feu de forêt sont le taux d'humidité ou encore la température. L'étude des régressions à aussi montrer que certaines variables pouvaient être liées comme par exemple DC avec DMC. Cependant, il faut toujours avoir à l'esprit que ce sont des analyses de données et des statistiques, ce n'est pas forcément la réalité.

Ce projet a été très intéressant tant d'un point de vue connaissance mais aussi d'un point de vue travail de groupe. Nous avons en effet pu mettre en application (en plus des travaux dirigés) notre cours sur un problème concret, cela a été très enrichissant. Le travail de groupe nous a appris à savoir nous organiser, en effet avec la crise sanitaire et la différence d'emplois du temps ce fut difficile de se retrouver, c'est pourquoi nous avons rapidement décidé de travailler à distance via Discord. Nous nous connections sur Discord sur un salon vocal afin de travailler sur le code mais aussi la rédaction, ce qui nous a permis d'être efficace.

De plus, nous avons écrit ce rapport en LaTeX, ce qui fut une difficulté au début mais une fois pris en main, cela s'est avéré très pratique. Nous avons donc appris à utiliser un nouvel outil : Overleaf. Cet outil permet de coder en LaTex en ligne.

Nous sommes donc très satisfaits de notre projet et sommes fiers des résultats que nous obtenons.

Section 6

Annexes

Indice forêt - météo (FWI) Structure					
Index	Bas	Modéré	Haut	Très Haut	Extrême
FFMC	0 - 81	81 - 88	88 - 90.5	90.5 - 92.4	92.5 +
DMC	0 - 13	13 - 28	28 - 42	42 - 63	63 +
DC	0 - 80	80 - 210	210 - 274	274 - 360	360 +
ISI	0 - 4	4 - 8	8 -11	11 - 19	19 +
BUI	0 -19	19 -34	34 - 54	54 - 77	77 +
FWI	0 - 5	5 - 14	14 - 21	21 - 33	33 +

Figure 19 – Tableau de valeurs des différents indices du FMI

```
n, p = data.shape
data_norm = (data - np.mean(data, axis=0)) / np.std(data, axis=0)
cor = 1/n * data_norm.T@data_norm
valeurP, vecteurP = np.linalg.eigh(cor)
print(valeurP)

#Ordre à modifier pour avoir les variances dans l'ordre décroissant
indexes_sorted = np.argsort(valeurP[::-1])
valeurP = valeurP[indexes_sorted]
print(valeurP)
vecteurP = vecteurP[:,indexes_sorted]
attributsSorted = attributs[indexes_sorted]

#Projection + reconstruction
P = vecteurP[:,:4]
Xproj = data_norm@P
Xrec = ((Xproj@P.T)*np.std(data, axis=0))+np.mean(data, axis=0)
```

FIGURE 20 – ACP