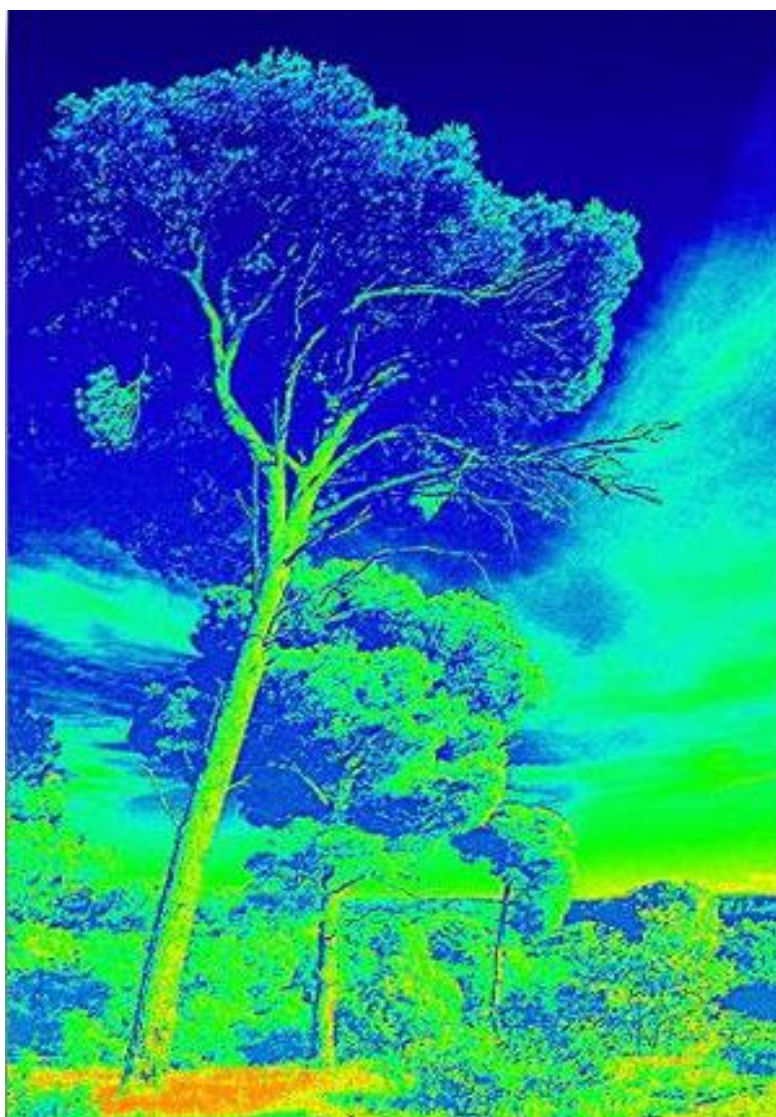


Elaboration de modèles prédictifs concernant les feux de forêts dans le sud de la France



Le 12/07/2017

RESUME

Les feux de forêts constituent dans la région méditerranéenne un phénomène ancien et récurrent. Ces feux ont, la plupart du temps, une origine humaine. Parmi ceux-ci, il faut toutefois distinguer les feux provoqués sciemment et de manière contrôlée des feux accidentels. La première catégorie inclut en particulier les feux dits de montagne, résultant de l'activité pastorale et apparaissant en dehors de la période estivale [1], tandis que les feux estivaux résultent en général de l'imprudence voire de la malveillance et provoquent beaucoup plus de dégâts.

Mieux anticiper ces derniers constitue un enjeu de taille, aussi bien sur le plan humain qu'écologique et économique. De ce fait, grâce à une meilleure compréhension du comportement des feux de forêt, il serait possible d'anticiper les risques de grands incendies et donc de déployer les moyens d'interventions de manière proportionnée au risque.

Ce rapport a donc pour objectif de modéliser l'extension des incendies grâce à un modèle prédictif faisant intervenir un certain nombre de paramètres, parmi lesquels les conditions météorologiques. Nous utiliserons notamment pour cela les bases de données Prométhée et weatherunderground. La première recense l'ensemble des feux de forêts dans 15 départements du quart Sud-Est de la France et la deuxième contient les historiques des bulletins météorologiques de nombreuses stations dans le monde.

Bien que les conditions météorologiques jouent un rôle évident dans l'apparition et le développement des feux, rien ne nous garantit toutefois qu'elles soient suffisantes pour élaborer un modèle fiable ; il nous faudra donc également vérifier s'il n'est pas nécessaire pour cela de leur adjoindre d'autres données.

REMERCIEMENTS

Je tiens premièrement à remercier toute l'équipe de formation de l'école centrale de Paris, aussi bien pour l'organisation de la formation que pour la qualité des enseignements fournis. Mes remerciements s'adressent en particulier à Madame Aufaure pour son encadrement ainsi que l'enseignement qu'elle nous a dispensé dans le domaine des bases de données, à Monsieur Guillemot pour son enseignement des probabilités et son initiation au langage R ainsi qu'à Monsieur Tenenhaus pour tous ses précieux cours de statistique. Dans chacun de ces domaines, mes connaissances préalables étaient relativement modestes voire nulles, mais la qualité et le dynamisme de leur enseignement m'a fait progresser très rapidement malgré la difficulté des sujets abordés. Plus généralement, la formation a suscité chez moi un grand engouement pour ces disciplines si bien que je continue encore aujourd'hui à approfondir les connaissances ainsi acquises.

J'adresse ma plus grande gratitude au personnel de la société Open. Je pense notamment à Mesdames Billy-Mauduit et Gorneau, en charge des formalités administratives et de la logistique de ma formation.

Je remercie également Monsieur Dumont, directeur de projet au sein de la business unit d'Open Rennes pour ses précieux conseils en cartographie et systèmes d'information géographique.

Je tiens enfin à remercier Monsieur Cabane, chargé de mission à la délégation à la protection de la forêt méditerranéenne pour sa promptitude à répondre à mes questions et pour l'intérêt qu'il porte à mes travaux.

TABLE DES MATIERES

1.1	Contexte.....	5
1.1.1	Présentation générale de la zone Prométhée	5
1.1.2	Découpage de la zone Prométhée	6
1.1.3	Caractérisation des feux de forêt.....	8
1.1.4	Organisation des services de lutte et de prévention des incendies.....	9
1.2	Enjeux.....	10
2	DONNEES DISPONIBLES	11
2.1	Base de données Prométhée	11
2.2	Extraction des bulletins météorologiques.....	12
2.3	Conclusion	13
3	PREPARATION DES DONNEES A ANALYSER	14
3.1	Présentation de MongoDB	14
3.2	Préparation d'un échantillon de données incendie et météo	15
3.3	Conclusion	17
4	ANALYSE DESCRIPTIVE ET PREMIERES TENDANCES	18
4.1	Evolution des feux de forêt depuis 1973.....	18
4.2	Analyse descriptive : exemple de la Haute-Corse	18
4.3	Conclusion	19
5	ELABORATION D'UN MODELE PREDICTIF	20
5.1	Préparation des données	20
5.2	Elaboration d'un modèle prédictif à partir des données météo.....	21
5.3	Prise en compte des données environnementales	22
5.3.1	Influence de la végétation	22
5.3.2	Influence des variables démographiques.....	22
5.3.3	Résultats	23
5.4	Prise en compte des données d'intervention	24
5.4.1	Corrélations entre les variables liées à l'intervention	24
5.4.2	Elaboration d'un modèle prédictif.....	25
5.5	Bilan	26
6	CONCLUSION	27

1 INTRODUCTION

1.1 Contexte

Chaque année, en moyenne, 24000 hectares de forêt sont brûlés en France métropolitaine. Ceci représente environ 0,15 % de la surface forestière française [2].

Les territoires français les plus touchés sont situés dans le quart sud-est de la France ainsi que dans la région Aquitaine, comme le montre la carte ci-dessous (extraite de [2]), qui représente l'ensemble des communes exposées aux risques de feux de forêt.



Fig. 1 : Cartographie des communes exposées au risque de feu de forêt.

Le risque d'incendie dans ces régions est en grande partie dû au climat aride qui leur est propre ainsi qu'à la nature de la flore fortement combustible qui constitue leurs forêts (conifères, garrigue, maquis,...). Nous pouvons toutefois noter que certaines régions du nord de la France (Bretagne, Centre, Pays de Loire) seront soumises aux mêmes risques d'ici 2040 du fait du réchauffement climatique [3].

1.1.1 Présentation générale de la zone Prométhée

Nous nous focaliserons dans la suite de cette étude sur les feux de forêts survenus au sein du quart sud-est de la France. Nous nous appuierons pour cela sur les données issues de la base Prométhée qui recense l'ensemble des feux de forêts survenus depuis 1973 dans les départements à risque suivants : Alpes de Haute Provence, Alpes Maritimes, Ardèche, Aude, Bouches du Rhône, Corse du Sud, Drôme, Gard, Haute Corse, Hautes Alpes, Hérault, Lozère, Pyrénées Orientales, Var et Vaucluse.

Dans la zone couverte par la base de données Prométhée, sur l'ensemble des années 2010 à 2014, 1355 feux de forêts ont été enregistrés en 2010, 1744 en 2011, 1857 en 2012, 1205 en 2013 et 1290 en 2014, soit une moyenne d'environ 1500 feux de forêt par an. La surface totale de forêt brûlée est d'environ 6200 hectares en 2010, 4500 en 2011, 4400 en 2012, 1920 en 2013 et 4100 en 2014, soit une moyenne d'environ 4200 hectares par an.

Afin de nuancer ces chiffres, il nous faut toutefois relever que de nombreux feux de forêt interviennent en dehors de la période estivale (du mois d'octobre au mois de juin) ; ces feux correspondent aux feux pastoraux, dits « feux de montagne » et sont souvent provoqués de manière contrôlée (et réglementée) par les éleveurs afin de faciliter la transhumance de leurs troupeaux [1]. Ainsi, en 2014, ils représentaient 40% des feux de forêt et 29% de la surface

brûlée. Les feux estivaux seront donc à considérer avec plus d'attention du fait des conditions météo défavorables et de leur origine souvent accidentelle.

La figure suivante, extraite de [2], illustre le nombre de départs de feux dans chacun des 15 départements de la base de données Prométhée de 1973 à 2009.

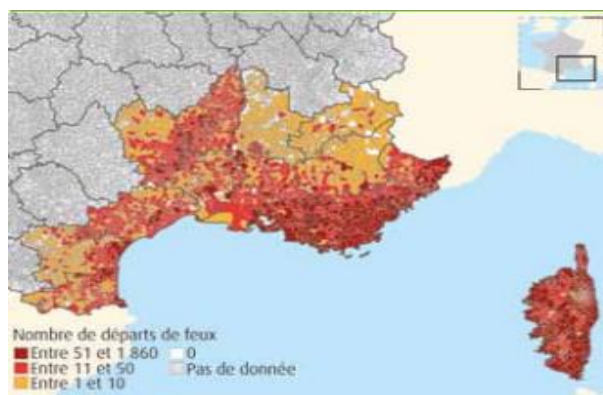


Fig. 2 : nombre de départs de feux par commune dans le sud-est de la France.

1.1.2 Découpage de la zone Prométhée

Comme nous venons de le mentionner, il est nécessaire de distinguer les feux estivaux des feux pastoraux. Ces derniers interviennent principalement dans les régions montagneuses et agricoles tandis que les premiers surviennent majoritairement dans les départements littoraux subissant une fréquentation estivale importante.

Ce chapitre a pour objectif de mettre en évidence ces disparités au sein de la zone Prométhée. L'ensemble du code associé à ce chapitre est présenté dans l'annexe A1. Les résultats ont été obtenus à partir des données issues de Prométhée de l'année 2010 à l'année 2014. Pour cela, les incendies ont été agrégés sur les 12 mois de l'année et les départements. Il en résulte le tableau suivant :

	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août
83	4.380952	3.428571	8.380952	5.523810	5.714286	7.619048	19.238095	23.80952
2A	1.945525	2.101167	8.093385	6.614786	7.937743	17.042802	16.420233	12.29572
26	1.098901	14.285714	10.989011	13.186813	9.890110	5.494505	16.483516	21.97802
30	0.660066	5.610561	6.930693	6.930693	8.250825	10.561056	24.422442	23.76238
84	2.884615	10.576923	6.730769	8.653846	10.576923	11.538462	19.230769	14.42308
04	3.428571	16.000000	20.000000	7.428571	1.714286	4.571429	19.428571	13.14286
06	7.396450	8.875740	16.272189	10.946746	10.650888	6.804734	7.692308	14.20118
11	1.854494	7.132668	9.985735	8.131241	5.563481	10.271041	16.547789	18.40228
2B	5.390836	1.415094	7.479784	2.628032	3.099730	6.671159	11.388140	16.64420

Les départements présentant un nombre total d'incendies plus ou moins élevé, chaque ligne du tableau est ensuite normalisée par ce nombre total. Un clustering hiérarchique appliqué à ce tableau donne le dendrogramme suivant ; puisque nous avons affaire à des valeurs numériques, la distance utilisée est la distance euclidienne classique et l'ultramétrie la distance de Ward.

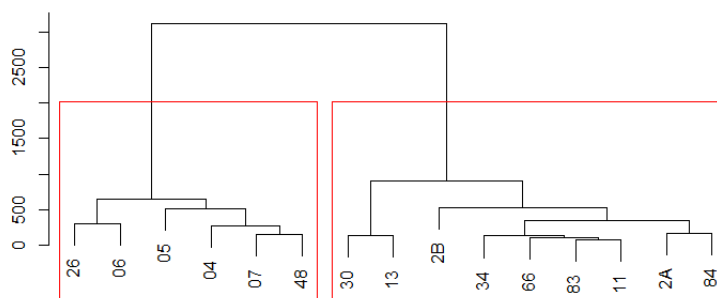


Figure 3 : dendrogramme résultant du clustering hiérarchique appliqué au tableau.

Nous distinguons effectivement 2 groupes de départements qui correspondent, hormis les Alpes Maritimes (département 06) à la distinction suivante :

- Départements littoraux : Var (83), Corse du sud (2B) et Haute Corse (2A), Bouches du Rhône (13), Gard (30), Hérault (34), Pyrénées Orientales (66), Aude (11), Vaucluse (84).
- Départements de l'intérieur : Lozère (48), Drôme (26), Hautes Alpes (07), Alpes de Haute Provence (05).

Cette tendance se confirme sur le bi-plot de la figure suivante, issu de l'analyse en composantes principales appliquée au même tableau de valeurs :

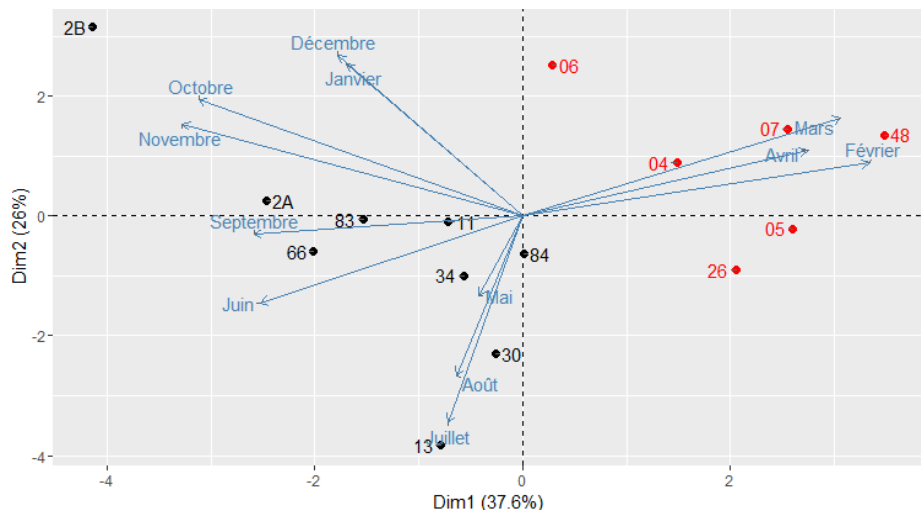


Figure 4 : bi-plot résultant de l'ACP.

Nous constatons tout d'abord que les 2 axes factoriels expliquent à eux deux plus de 64% de la variance totale. Nous observons de nouveau le regroupement des départements selon 2 grandes classes : les départements littoraux, fortement corrélés à la saison estivale (de mai à septembre) et les départements de l'intérieur, fortement corrélés à la saison hivernale (février, mars). Nous pouvons toutefois remarquer le cas particulier de la Corse du Sud (2B), qui a pour particularité de présenter, parmi les départements littoraux, en plus de la période estivale, de nombreux feux de forêt au mois d'octobre (20%), soit une quantité proche des mois de juillet et août.

Nous avons donc finalement regroupé les départements de la zone Prométhée en 3 catégories : les départements littoraux (sans les départements corses), les deux départements corses et les départements de l'intérieur. Les graphiques des figures 5a, 5b et 5c illustrent respectivement, pour chacun de ces 3 groupes, la répartition de la proportion des incendies pour chaque département selon les 12 mois de l'année.

Nous observons bien sur la figure 5a que la partie littorale se caractérise par un pic du nombre d'incendies en période estivale (mois de juillet et août), avec une prédominance des Bouches du Rhône et du Var.

Dans le cas particulier de la Corse (figure 5b), la même tendance est observée avec, toutefois, une prolongation du pic estival sur les mois de septembre et octobre.

Enfin, les départements de l'intérieur (figure 5c), plus montagneux, présentent un pic hivernal au mois de mars qui précède un second pic estival moins important réparti sur les mois de juillet et août, avec une prédominance des Alpes Maritimes et de l'Ardèche.

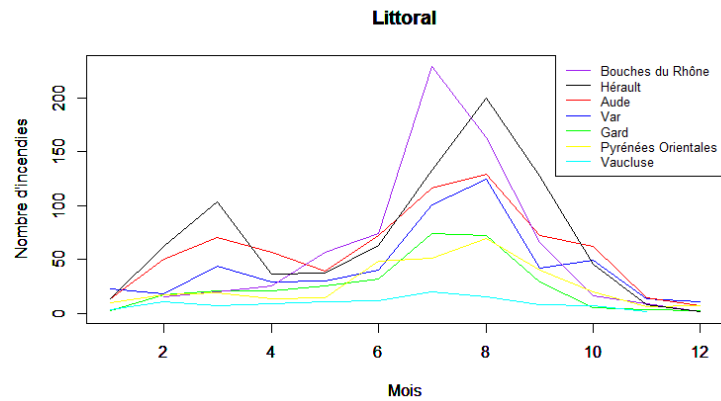


Figure 5a : répartition de la proportion des incendies selon les 12 mois de l'année pour chaque département littoral de la zone Prométhée.

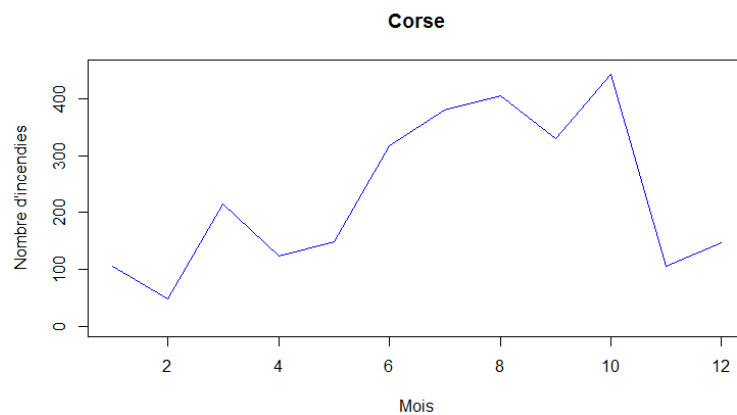


Figure 5b : répartition de la proportion des incendies selon les 12 mois de l'année pour l'ensemble de la Corse.

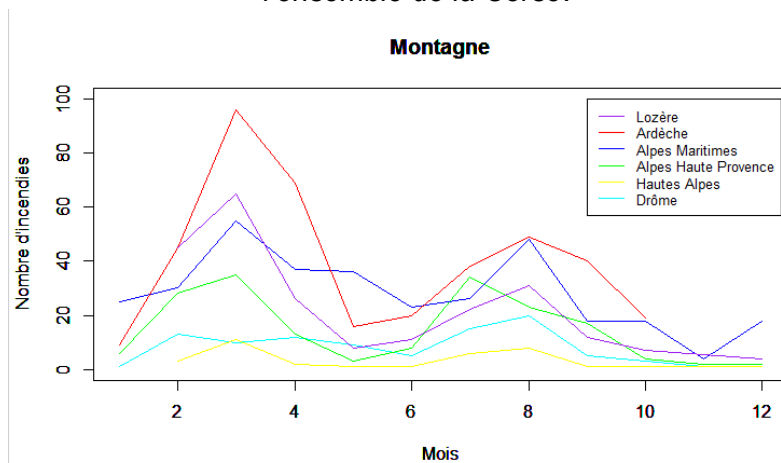


Figure 5c : répartition de la proportion des incendies selon les 12 mois de l'année pour chaque département de l'intérieur de la zone Prométhée.

1.1.3 Caractérisation des feux de forêt

Une grande partie des feux de forêts recensés (estivaux ou non) sont à classer parmi les petits feux de forêt, c'est-à-dire lorsque la surface brûlée est inférieure à 1 hectare [2] ; dans le cas contraire, nous parlons alors d'incendie de forêt. Ainsi, en 2010, 1092 feux de forêt s'étendaient sur moins d'un hectare, 1415 en 2011, 1459 en 2012, 1004 en 2013 et 1007 en

2014. La proportion des petits feux de forêts varie donc peu d'une année sur l'autre et se trouve être de l'ordre de 80%.

A contrario, les grands feux de forêts estivaux ayant dévasté plus de 100 hectares étaient au nombre de 5 en 2010, 9 en 2011, 5 en 2012, 2 en 2013 et 6 en 2014. Ils représentent donc une infime minorité parmi l'ensemble des feux de forêt mais sont à l'origine de 4300 hectares de forêt brûlés en 2010 (soit 70% de la surface brûlée cette année-là), 1518 hectares en 2011 (30%), 1303 hectares en 2012 (30%), 1300 hectares en 2013 (22%) et 1711 hectares en 2014 (41%).

En revanche, bien qu'ils soient majoritaires en nombre, les feux de forêts inférieurs à 1 hectare ne contribuent à la surface brûlée totale qu'à hauteur de 3% en 2010, 5% en 2011, 6% en 2012, 8% en 2013 et 4% en 2014. Leur impact est donc négligeable.

L'étude de la répartition des incendies suivant la surface brûlée sera reprise dans la partie 4.

1.1.4 Organisation des services de lutte et de prévention des incendies

Pour lutter efficacement contre les incendies dans la zone Prométhée, des moyens matériels conséquents ont été mis en œuvre. Environ 700 sapeurs-pompiers et 700 membres des unités d'interventions civiles sont chargés des interventions au sol, tandis que les interventions aériennes sont assurées par 23 aéronefs bombardiers d'eau ainsi que 3 avions de reconnaissance [2].

La politique de déploiement de ces moyens s'organise quant à elle au niveau départemental par l'intermédiaire des SDIS (Services Départementaux d'incendie et de Secours) soumis à l'autorité du préfet. Ceci explique en partie des disparités relativement importantes concernant l'ampleur des feux de forêt d'un département à l'autre. L'histogramme de la figure 6 représente la surface brûlée moyenne par incendie pour chaque département, cette surface étant calculée sur l'ensemble des incendies apparus entre 2002 et 2015.

Tout d'abord, nous constatons que, bien qu'ils présentent le plus grand nombre d'incendies annuels (cf. figure 7), le Var et la Corse du Sud sont les départements pour lesquels les incendies sont les moins dévastateurs. Par contre, nous observons que le département des Hautes-Alpes, qui compte très peu d'incendies annuels (nombre d'incendies inférieur à 10 par an d'après la figure 7) se situe en troisième position concernant l'ampleur des incendies.

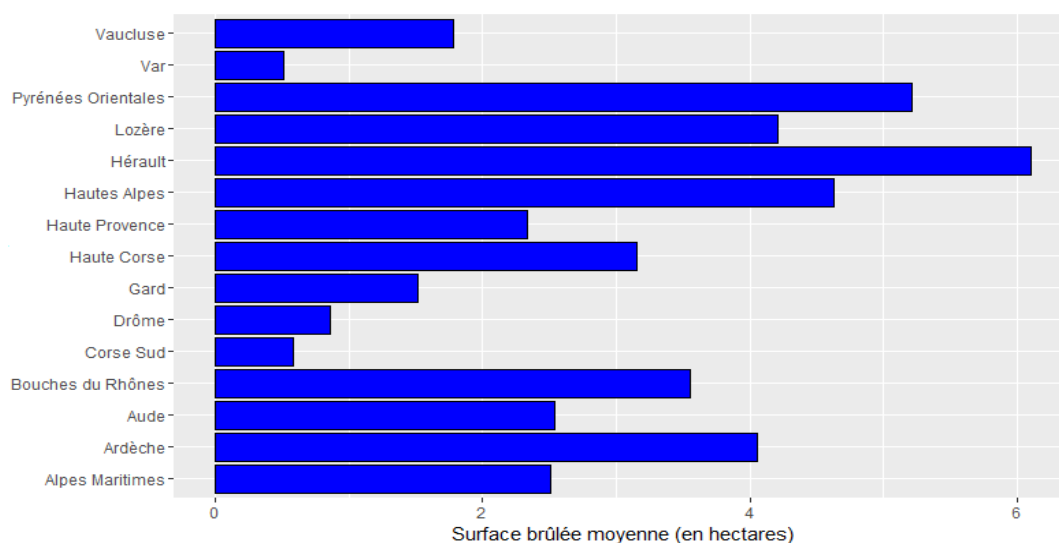


Fig.6 : surface moyenne brûlée par incendie pour chaque département.

Nous pouvons en conclure que l'ampleur et le nombre d'incendies sont totalement décorrélés. L'ampleur moyenne des incendies dépendra en effet fortement des moyens d'intervention propres à chaque département ainsi que de la topographie du terrain. Nous pouvons en effet supposer qu'un département fortement montagneux, à l'instar des Hautes Alpes rendra en général plus difficile l'intervention des moyens de secours. A contrario, les départements littoraux présentent un nombre plus important d'incendies du fait de la nature de leur végétation et de leur démographie plus importante (rappelons que les incendies sont la plupart du temps d'origine humaine) mais sont de ce fait mieux préparés à intervenir, d'où la plus faible ampleur de leurs incendies.

Le code associé à l'histogramme est présenté en annexe A2.

1.2 Enjeux

Sur le plan humain, bien que les décès liés aux feux de forêts soient très peu nombreux en France, les activités humaines telles que les zones agricoles ou industrielles ainsi que les réseaux de communication sont régulièrement touchés. D'un point de vue écologique, bien que certaines espèces telles que les pins d'Alep se régénèrent très rapidement, les grands incendies détériorent considérablement les sols via la perte des minéraux et la disparition de la couverture végétale, ce qui peut entraîner des phénomènes d'érosion [2]. La faune est également touchée, par exemple la tortue de Hermann, espèce rare vivant seulement dans le sud de la France dont l'incendie de Montfort-Correns a détruit l'habitat. Enfin, sur le plan économique, le coût de la protection des forêts est évalué à plus de 500 millions d'euros par an [2,3].

De plus, comme nous l'avons mentionné plus haut, l'ensemble de ces coûts risque de s'accroître durant les prochaines décennies du fait du réchauffement climatique. L'augmentation de la quantité de surface brûlée annuelle est ainsi estimée à 140% à l'horizon de l'année 2070 [3].

Puisqu'il est évidemment difficile voire impossible de prévoir la localisation exacte d'un départ de feu, l'enjeu majeur de la lutte contre les incendies réside dans l'optimisation de l'organisation spatiale et temporelle des moyens d'intervention, c'est-à-dire faire en sorte qu'un maximum de moyens soit disposé là où le risque de feux de forêts est le plus important et qu'ils puissent intervenir le plus rapidement possible afin de limiter leur propagation. A titre indicatif, en présence de vents forts, un feu de forêt peut se propager à une vitesse de l'ordre de 5 km/h [2]. Les avions de reconnaissance et les canadiens sont les plus concernés par cette problématique car ils sont rapidement déplaçables d'un département à l'autre

Nous proposons donc dans ce rapport d'anticiper le comportement des feux de forêt grâce à la mise en œuvre d'un modèle prédictif. Comme tout modèle prédictif, il nous faudra nous appuyer sur un ensemble de paramètres susceptibles d'avoir un impact sur le développement d'un incendie, parmi lesquels les conditions météorologiques.

2 DONNEES DISPONIBLES

Nous présentons dans cette partie les bases de données Prométhée et wunderground qui nous fourniront par la suite, respectivement, les informations liées aux feux de forêt et les bulletins météorologiques associés.

2.1 Base de données Prométhée

La base de données Prométhée recense tous les feux de forêts survenus depuis 1973 dans les 15 départements du sud-est de la France suivants : Alpes de Haute Provence, Alpes Maritimes, Ardèche, Aude, Bouches du Rhône, Corse du Sud, Drôme, Gard, Haute Corse, Hautes Alpes, Hérault, Lozère, Pyrénées Orientales, Var et Vaucluse. Ils sont récapitulés sur la carte ci-dessous, extraite du site www.promethee.com :

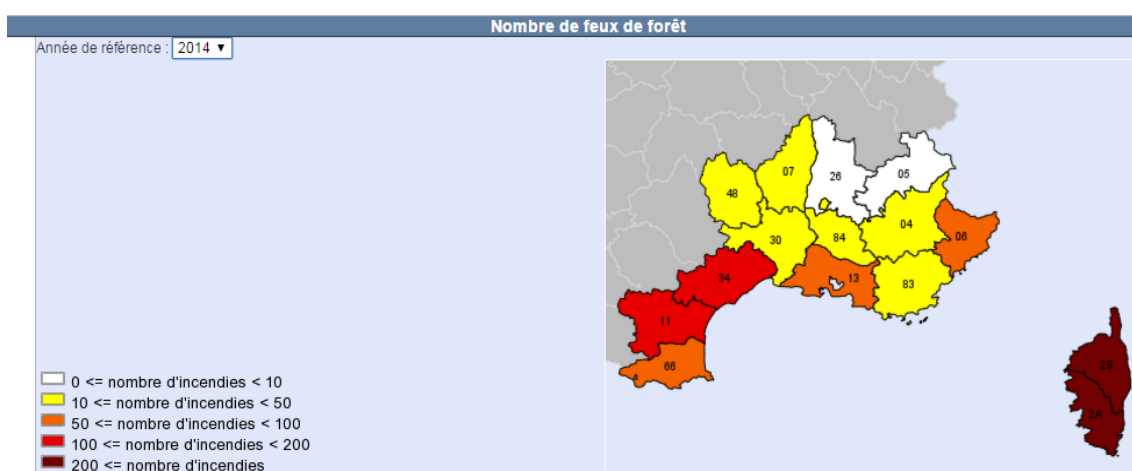


Fig. 7 : les 15 départements français couverts par la base de données Prométhée.

Les requêtes au sein de la base sont effectuées sur la page du site suivante :

Critères de sélection

Type de feu : Forêt ?

Date : du 01/01/2014 ? au 31/12/2014 ?

Horaires : de h à h ?

Surface : de ha à ha ?

Nature de la cause : -

Origine de l'alerte : -

Type de dommage : -

Toute la zone Prométhée ☒

Région administrative : -

Département : -

Commune / code INSEE : ?

Code du carreau DFCI : -

Coordonnées géographiques : -

Sélectionner les incendies Sélection par défaut

Carte

Tableau










	Num.	Type	Alerte↑	Dpt.	Commune	Surface(ha)	Nature de la cause
	50	Forêt	02/01/2014	66	Salses-le-Château	1,0000	Travaux (Particuliers)
	14	Forêt	07/01/2014	06	Lantosque	4,8400	-
	1559	Forêt	09/01/2014	2A	Lecci	0,0010	-
	6744	Forêt	11/01/2014	2B	Canari	0,0100	-
	106	Forêt	13/01/2014	30	Branoux-les-Taillades	0,2500	-
	51	Forêt	17/01/2014	66	Sorède	1,0000	Travaux Forestiers
	47	Forêt	18/01/2014	2B	Poggio-Mezzana	0,3000	-
	105	Forêt	26/01/2014	06	Saint-Jeannet	0,3000	Intérêt
	40	Forêt	26/01/2014	2B	Bigorno	1,0000	-

Fig. 8 : page de sélection des feux de forêt.

Comme nous pouvons le constater, il est possible d'appliquer différents filtres à la sélection : la période qui nous intéresse, la cause de l'incendie, le département,... Enfin, le tableau résultant peut être exporté dans un fichier au format CSV.

2.2 Extraction des bulletins météorologiques

Les conditions météorologiques régnant lors de chaque feu de forêt seront extraites du site www.wunderground.com. Ce site recueille les bulletins de nombreux services nationaux mais reçoit également en temps réel les données des stations météorologiques de 180000 contributeurs anonymes à travers le monde. Il fournit également des images satellites et radar de l'atmosphère ainsi que des cartes d'alertes météo (tempêtes, fortes chaleurs,...) en Europe et aux Etats-Unis. De plus, certains bulletins sont archivés et il est ainsi possible de remonter jusqu'à l'année 1996 en ce qui concerne les bulletins français. La figure suivante nous fournit l'exemple du bulletin du 8 septembre 2011 associé à la ville d'Antibes :

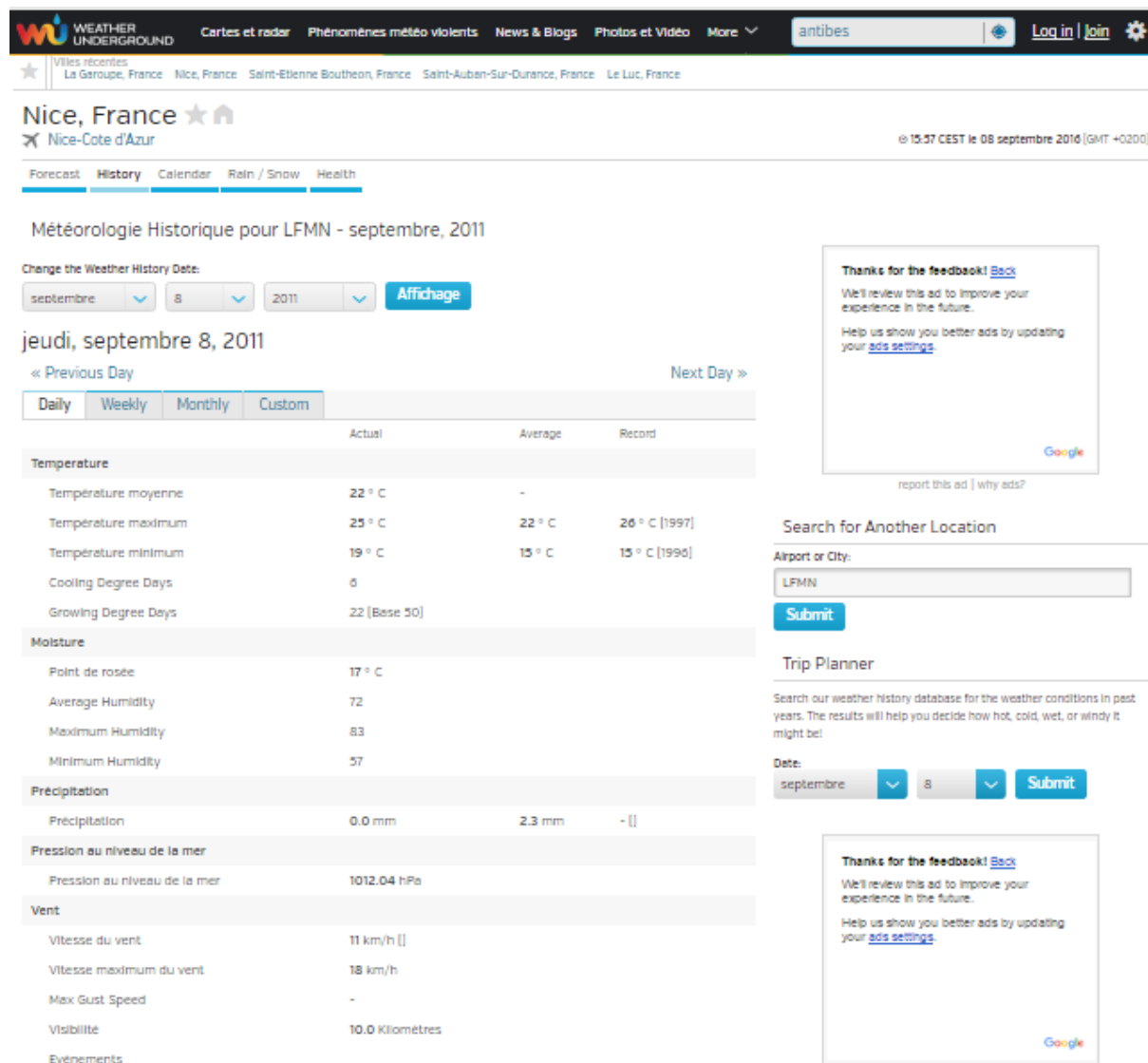


Fig. 9 : exemple d'archive de bulletin météorologique

Nous pouvons observer que la station météorologique retenue est celle de l'aéroport de Nice. Ceci est dû au fait que le site n'archive en fait que les bulletins des aéroports civils et militaires. C'est donc l'historique de l'aéroport le plus proche du lieu demandé qui est retenu lors de la requête. L'ensemble des informations qui nous intéressent sont toutefois toujours disponibles : température, humidité, taux de précipitation, vent,...

Nous avons positionné sur les cartes des figures 10a et 10b l'ensemble des aéroports du quart sud-est de la France dont les historiques météorologiques sont accessibles sur le site.

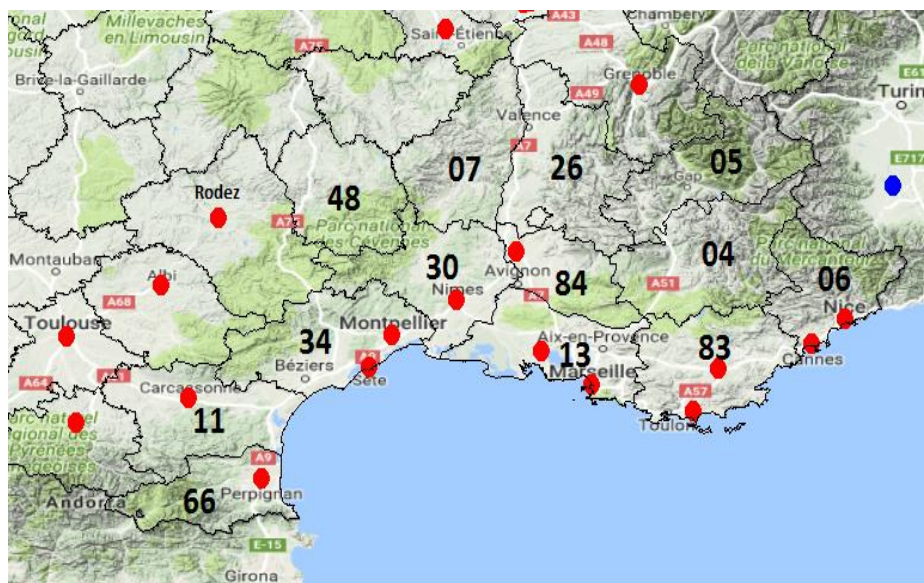


Fig. 10a



Fig. 10b

Figures 10a et 10b : cartographie des aéroports et bases aériennes du sud-est de la France métropolitaine (10a) et de la Corse (10b) dont l'historique météorologique est disponible.

Evidemment, les aéroports ne sont pas répartis de manière uniforme au sein du réseau Prométhée. Nous pouvons ainsi remarquer sur ces deux figures que 5 départements n'ont pas d'historique météorologique sur leur territoire : la Lozère (48), l'Ardèche (07), la Drôme (26), les Hautes Alpes (05) et les Alpes de Haute Provence (04). Ainsi, une requête concernant un incendie survenu en Lozère nous renverra par exemple vers les archives météo des aéroports de Rodez ou Nîmes. L'imprécision des données sera donc plus importante dans ces départements. Notons que nous avons retenu la possibilité de requêter les archives de l'aéroport italien de Levaldigi (point bleu sur la figure 10a) afin de désenclaver certaines communes proches de la frontière italienne et éloignées des aéroports français, les deux aéroports les plus proches pour cette région géographique étant ceux de Nice et Grenoble.

Le code R à l'origine des figures 10a et 10b est présenté en annexe B.

2.3 Conclusion

Nous avons présenté dans cette partie les deux sources de données sur lesquelles nous nous appuyons dans la suite. Lors de ces travaux, nous aurons notamment à retrouver sur le site wunderground.com les conditions météorologiques concomitantes à chaque incendie. Malheureusement, nous avons pu mettre en évidence dans cette partie que les historiques météo ne sont disponibles sur ce site que pour les stations des aéroports. C'est donc l'aéroport le plus proche du lieu de l'incendie qui sera retenu à chaque fois. Cette limitation est évidemment source d'imprécisions, en particulier dans certains départements montagneux où les aéroports sont moins nombreux, voire absents. Il nous faudra dans ce cas envisager des requêtes vers des aéroports situés dans des départements limitrophes.

3 PREPARATION DES DONNEES A ANALYSER

Nous avons présenté dans la partie précédente les sources qui nous fourniront les données indispensables pour la suite de nos travaux. Toutefois, avant de mettre en place un modèle prédictif, il peut s'avérer utile de réaliser une première analyse sur un nombre réduit d'échantillons. Nous pouvons ainsi obtenir rapidement et simplement un ensemble de tendances (définition de seuils, repérage de valeurs aberrantes, identification d'un comportement général...) susceptibles de nous servir de fil conducteur pour l'élaboration d'un modèle plus complexe fiable et robuste.

De plus, comme nous le verrons dans la partie 5, nos travaux nécessitent, pour chaque incendie, d'accéder au site wunderground.com via une URL définissant notamment sa date et l'aéroport le plus proche. Ensuite, une procédure de webscrapping est mise en œuvre afin de récupérer automatiquement les informations météo sur la page internet associée à l'URL. Tout ceci s'avère relativement coûteux en terme de temps d'exécution, étant donnée la quantité d'incendies répertoriés. De ce fait, dans le cadre d'une première estimation, nous pouvons par exemple restreindre les incendies à un seul département sur une ou plusieurs années et extraire depuis wunderground.com l'historique météorologique d'un aéroport de ce département sur la même période au format CSV. Nous évitons ainsi l'étape de webscrapping et l'analyse des données est quasi immédiate.

Nous allons donc montrer dans cette partie comment nous avons utilisé le SGBD orienté document MongoDB pour extraire un ou plusieurs échantillons pertinents à partir des données incendie et météo au format CSV. Nous présentons brièvement dans le chapitre suivant MongoDB ainsi que l'IHM associée nommée Robomongo.

3.1 Présentation de MongoDB

Le logiciel Robomongo, présenté sur la figure 11, offre une interface permettant d'interagir avec des bases de données MongoDB. Il est possible par son intermédiaire de visualiser les bases de données, les collections qu'elles contiennent, d'exécuter des scripts sur celles-ci,... Dans la terminologie de MongoDB, un ensemble d'enregistrements est stocké au sein d'une structure appelée collection. Les collections sont elles-mêmes regroupées dans des bases de données. Chaque champ d'une collection est nommé document.

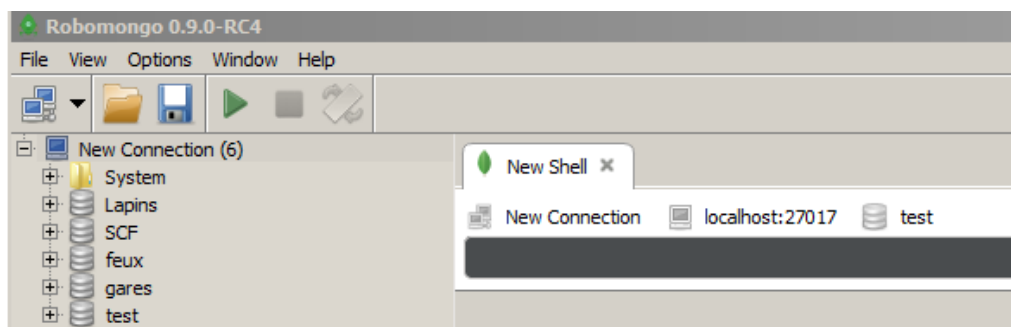


Figure 11 : présentation de l'interface client de MongoDB

Le menu présent dans la partie gauche de la fenêtre répertorie l'ensemble des bases de données présentes sur un serveur donné. La partie droite de la fenêtre permet quant à elle de lancer les requêtes sur les bases.

Comme nous pouvons le constater, nous avons créé la base de données **feux**. Nous allons donc maintenant pouvoir y insérer des collections de documents.

3.2 Préparation d'un échantillon de données incendie et météo

Pour cela, nous avons extrait du site Prométhée le relevé des incendies au format CSV s'étalant sur 5 années, de 2010 à 2014. Pour intégrer une nouvelle collection dans la base **feux** à partir de ce fichier, nous avons eu recours au programme *mongoimport* qui se lance depuis l'invite de commande. La ligne de commande associée est la suivante:

```
mongoimport --db feux --collection feux --headerline --type csv --file C:\feux.csv
```

Le paramètre *db* définit la base de données devant contenir la nouvelle collection, *collection* définit le nom de la collection, *headerline* permet de choisir si les noms des champs de la collection doivent être définis à partir de la première ligne du fichier tandis que le paramètre *file* correspond au fichier CSV à lire.

La collection **feux** apparaît maintenant au sein de la base de données **feux** :

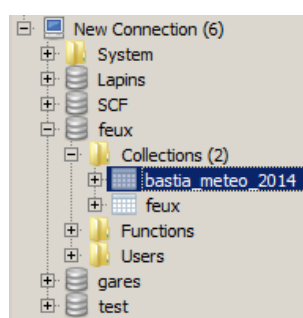


Fig. 12 : création d'une nouvelle collection au sein de la base de données « feux ».

Nous souhaitons maintenant extraire un échantillon pertinent parmi ces données. D'après [2], la Corse est la région française la plus touchée, comme le confirme la carte de la figure 13 (extraite du site Prométhée) qui illustre les statistiques de l'année 2014 concernant le nombre de feux de forêt annuel. Par contre, la surface de forêt brûlée moyenne par incendie est beaucoup plus forte en Haute Corse qu'en Corse du Sud (cf. fig. 6), ce qui entraîne une superficie annuelle brûlée d'autant plus importante et rend les statistiques de ce département particulièrement intéressantes à étudier.

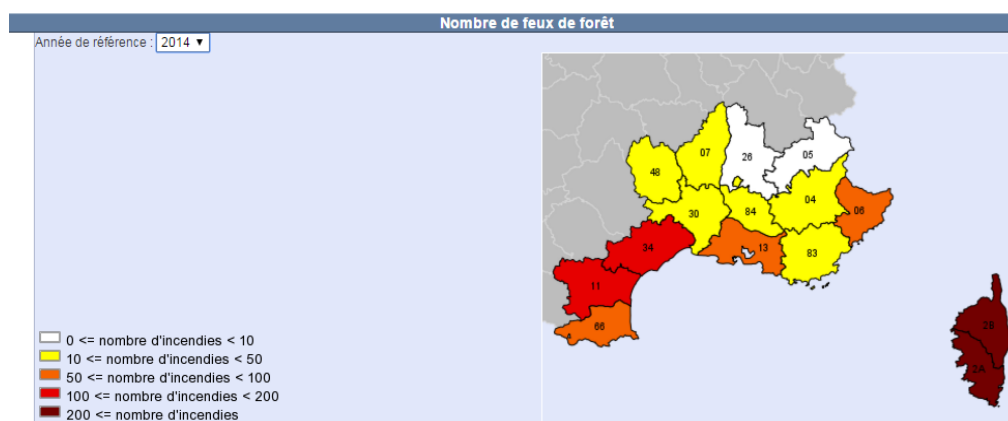


Fig. 13 : cartographie des feux de forêts survenus dans la zone Prométhée en 2014.

Nous choisissons donc de ne retenir pour notre premier échantillon que les feux de forêts survenus en Haute Corse. Nous n'avons alors qu'un seul relevé météorologique au format CSV à extraire du site *wunderground.com*, par exemple celui de l'aéroport de Bastia.

Nous pouvons ensuite restreindre notre échantillon à une année en particulier, par exemple l'année durant laquelle les feux de forêt ont été les plus nombreux en Haute Corse. Pour cela, nous avons exécuté le script suivant sur la collection **feux** :

```
1 use feux
2 // On va calculer pour chaque année la surface totale de forêt brûlée en Haute-Corse
3 // On utilise pour cela par exemple l'algorithme Map-Reduce
4 // La fonction mapFunction renvoie un couple clé-valeur à partir de chaque document
5 // La clé correspond à l'année et la valeur à la surface brûlée du document
6 var mapFunction = function() { emit(this.Annee, this.Surface); };
7 // La fonction reduceFunction réalise la somme des valeurs sur chaque clé
8 var reduceFunction = function(keyAnnee, valuesSurface) { return Array.sum(valuesSurface); };
9 // Application de l'algorithme à la collection feux.
10 // le paramètre "out" correspond à la collection en sortie
11 // le paramètre "query" limite le champ d'application au département de la Haute-Corse.
12 db.feux.mapReduce(mapFunction,
13                  reduceFunction,
14                  {out:"surface_brulee_annuelle",query:{Departement:"2B"}}
15                  )
16 // On affiche le contenu de la nouvelle base ainsi obtenue
17 db.surface_brulee_annuelle.find()
```

La ligne 1 indique que l'on utilise la base de données **feux**. Ensuite, l'algorithme Map-Reduce est mis en œuvre afin de compter le nombre total d'incendies pour chaque année. La collection **surface_brulee_annuelle** est produite en sortie. Nous constatons alors que la surface de forêt brûlée est la plus importante en 2014, suivie par 2011.

	_id	value
1	2010	2829260
2	2011	11879744
3	2012	4583884
4	2013	6494932
5	2014	20846242

Fig. 14 : surface brûlée annuelle entre 2010 et 2014 (en m²).

Par contre, si l'on s'intéresse au nombre de feux de forêt, nous constatons que l'année 2011 (396 feux) en compte plus que l'année 2014 (322 feux). Le script MongoDB ayant abouti à ce résultat est le suivant :

```
// On compte ici le nombre total d'incendies survenus en Haute-Corse pour chaque année
db.feux.find({Departement:"2B",Annee:2014}).sort({Commune:1}).count()
db.feux.find({Departement:"2B",Annee:2013}).sort({Commune:1}).count()
db.feux.find({Departement:"2B",Annee:2012}).sort({Commune:1}).count()
db.feux.find({Departement:"2B",Annee:2011}).sort({Commune:1}).count()
db.feux.find({Departement:"2B",Annee:2010}).sort({Commune:1}).count()
```

Les années 2011 et 2014 semblent donc être d'un intérêt particulier. En effet, bien que l'année 2011 présente un nombre de feux de forêts plus importants, la surface de forêt brûlée totale est quasiment double en 2014.

Le script suivant nous permet enfin d'extraire uniquement les feux de forêt survenus en Haute Corse durant l'année 2014. Le résultat du filtre est stocké dans la collection **feux_corse_2014**, créée au préalable via le logiciel Robomongo.


```
// Après avoir créé la base de donnée feux_corse_2014, on insère
// alors dans celle-ci chaque incendie survenu en Haute Corse en 2014
db.feux.find({Departement:"2B",Annee:2014}).forEach( function(x)
{db.feux_corse_2014.insert(x)} |)
```

La dernière étape consiste ensuite à fusionner cette dernière collection avec les données météo associées à l'aéroport de Bastia. Après avoir extrait au format CSV sur le site www.wunderground.com le relevé météo de cet aéroport sur l'ensemble de l'année 2014, nous l'avons importé au sein de la collection **bastia_meteo_2014** dans la base de données **feux** via *mongoimport*.

Nous avons ensuite mis en œuvre sur ces deux collections le script suivant :

```
// Regroupement des informations météo et incendies : on parcourt tous les documents de la collection feux_corse_2014
db.feux_corse_2014.find().forEach( function(e) {
// Pour chaque document, on sépare la chaîne de caractère du champ "Alerte" de part et d'autre du caractère " "
// La date est en effet écrite dans ce champ au format JJ/MM/YYYY HH:MM:SS
date = e.Alerte.split(" ");
// date est un tableau dont le premier élément est la date au format JJ/MM/YYYY, cohérent avec les dates
// du champ CET de la collection bastia_meteo_2014
// On recherche alors le document de la collection bastia_meteo dont le champ CET correspond à cette date
var res = db.bastia_meteo_2014.find({"CET":date[0]});
// On ajoute alors à chaque document de la collection feux_corse_2014 les champs suivants, issus du document ainsi trouvé :
// - Température max, moyenne et min
// - Humidité max, moyenne et min
// - Taux de précipitation max, moyen et min
// - Vitesse du vent max et moyenne
db.feux_corse_2014.update({_id: e._id}, {$set: {Tmax:res[0].Tmax , Tmoy:res[0].Tmoy , Tmin:res[0].Tmin,
Hmax:res[0].Hmax , Hmoy:res[0].Hoy , Hmin:res[0].Hmin,
Pmax:res[0].Pmax , Pmoy:res[0].Pmoy , Pmin:res[0].Pmin,
Vmax:res[0].Vmax , Vmoy:res[0].Vmoy} }) } |);
```

Pour chaque incendie de la collection **feux_corse_2014**, il récupère les données météo associées à partir de la date parmi la collection **bastia_meteo_2014** puis ajoute ces dernières au document concerné de la collection **feux_corse_2014**. Les données météo retenues sont : la température, le taux d'humidité, le taux de précipitation et la vitesse du vent (valeur minimum, moyenne et maximum pour chacune d'entre elles).

La collection **feux_corse_2014** est finalement exportée avec les données météo au format CSV via le programme *mongoexport* appelé depuis la ligne de commande suivante :

```
mongoexport --db feux --collection feux_corse_2014 --type csv --out C:\feux_corse_2014.csv --fields
Alerte,Surface, Tmax,Tmoy,Tmin,Hmax,Hmoy,Hmin, Pmax,Pmoy,Pmin,Vmax,Vmoy
```

3.3 Conclusion

Nous avons dans cette partie mis en œuvre le SGBD MongoDB afin de préparer des échantillons de données pertinents qui nous permettront dans la partie suivante de dégager des tendances générales concernant le comportement des feux de forêts en fonction de la météorologie. Les données de départ concernaient les années 2010 à 2014. Nous avons finalement retenu l'année 2014 dans la mesure où elle présente la plus grande surface de forêt brûlée.

4 ANALYSE DESCRIPTIVE ET PREMIERES TENDANCES

Cette partie a pour but de dégager quelques tendances statistiques concernant les feux de forêt en France. Nous pourrions ainsi justifier à partir de celle-ci certains choix opérés lors de l'élaboration des modèles prédictifs dans la partie suivante. Dans le premier chapitre, nous analyserons l'évolution générale de l'importance des feux de forêt de 1973 à nos jours. Dans le second chapitre, nous analyserons les échantillons concernant la Haute-Corse dont nous avons décrit l'extraction sur l'année 2014 dans la partie précédente.

4.1 Evolution des feux de forêt depuis 1973

Les années 90 ont marqué un tournant dans la lutte contre les feux de forêt [2]. En effet, la surface de forêt brûlée annuelle a diminué de moitié depuis 1990 (20000 hectares brûlés en moyenne par an depuis 1990 contre 37000 avant). Le nombre de feux de forêt annuels connaît également une tendance à la baisse, mais celle-ci reste néanmoins beaucoup moins marquée que la surface brûlée annuelle.

Les deux histogrammes suivants confirment ces deux tendances. Nous les avons obtenus à partir de la base de données Prométhée.

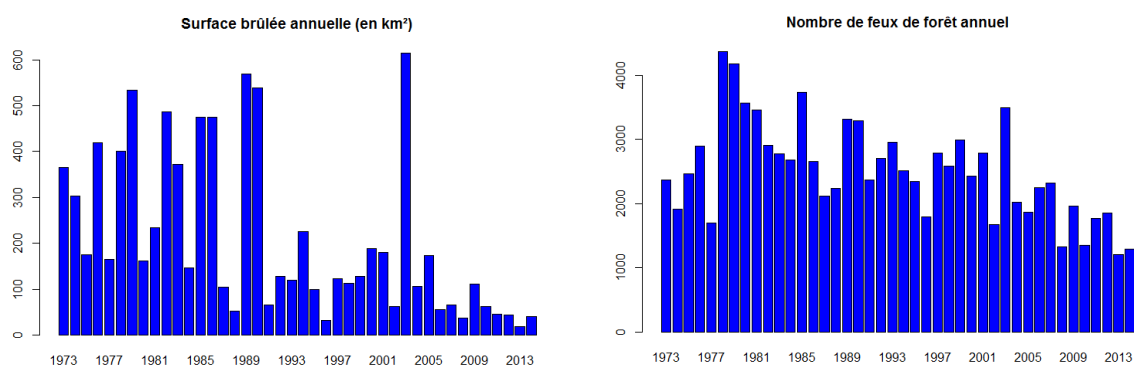


Figure 15a et 15b : respectivement, surface brûlée annuelle (en km²) et nombre de feux de forêt annuels de 1973 à 2014.

Il est donc impératif d'élaborer notre modèle à partir de données ultérieures à 1990 afin d'établir un modèle prédictif en phase avec les moyens d'intervention actuels.

4.2 Analyse descriptive : exemple de la Haute-Corse

Nous restreignons ici notre analyse au seul département de Haute Corse. Nous pouvons ainsi, en utilisant l'historique météorologique du département, recueillir rapidement des tendances pertinentes concernant l'influence de la météorologie sur les feux de forêt. Nous avons retenu l'année 2014 dans la mesure où celle-ci présente, sur les six dernières années, la superficie de forêt brûlée la plus importante.

L'histogramme de la figure 16 illustre la répartition des feux de forêt en fonction de la superficie brûlée. Comme nous pouvons le constater, une très large majorité de feux de forêt se sont répandus sur moins d'un hectare, ce qui justifie effectivement la classification de l'ONF (cf. 1.1) distinguant feux et incendies de forêt. Nous utiliserons donc également ce critère de discrimination dans la suite.

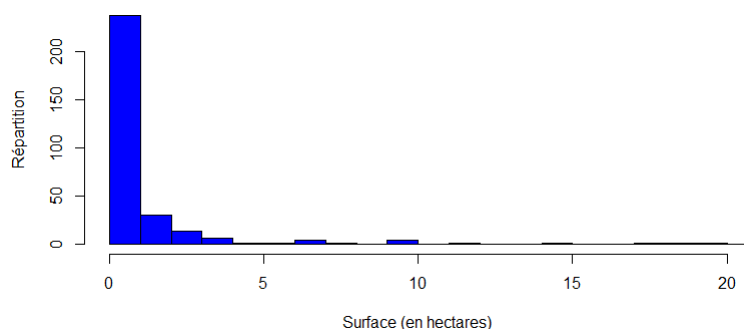
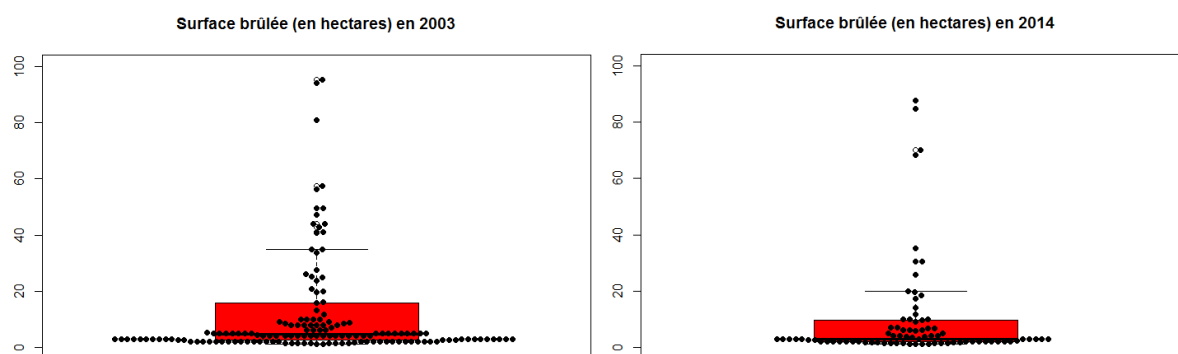


Figure 16 : répartition des feux de forêt en fonction de la superficie brûlée.

Les deux box-plot suivantes illustrent la surface brûlée par les incendies (surface brûlée supérieure à 1 hectare) survenus en 2003 et 2014. Nous comparons ces deux années afin de mettre en évidence l'influence de la température sur le développement des incendies. En effet, d'après la figure 15a, l'année 2003 constitue une année exceptionnelle en terme de surface de forêt brûlée parmi les années postérieures à 1990 ; ceci n'est pas un hasard dans la mesure où l'année 2003 a enregistré des records de température.



Figures 17a et 17b : répartition des incendies, respectivement en 2003 et 2014.

Pour une meilleure lisibilité, nous fixons la borne max à 100 hectares. A titre indicatif, les incendies ayant brûlé plus de 100 hectares sont au nombre de 15 en 2003 et 3 en 2014, l'incendie le plus dévastateur brûlant environ 5000 hectares en 2003 et 550 hectares en 2014.

Nous relevons un total de 85 incendies de forêt en 2014 et 164 en 2003 soit environ le double. De même, la hauteur du deuxième (≈ 10 hectares) et du troisième quartile (≈ 20 hectares) en 2014 est multipliée par 2 en 2003.

Les fortes températures (nous obtenons une température moyenne les jours d'incendies de l'ordre de 23°C en 2014 et $27,5^{\circ}\text{C}$ en 2003) favorisent donc à la fois la fréquence des incendies de forêt ainsi que leur propagation.

4.3 Conclusion

L'étude de l'évolution globale des feux de forêt depuis 1973 montre la nécessité d'utiliser des échantillons issus d'années postérieures à 1990 afin prendre en compte les progrès réalisés dans la lutte contre les incendies. L'étude de la répartition des feux de forêt selon la surface brûlée nous a ensuite conduits à conserver le seuil d'un hectare pour distinguer les « petits feux » des incendies de forêt. Enfin, nous avons mis en évidence la forte influence de la température sur la fréquence d'apparition des incendies de forêt ainsi que sur leur propagation.

Le code R associé aux chapitres 4.1 et 4.2 est exposé respectivement dans les annexes C1 et C2.

5 ELABORATION D'UN MODELE PREDICTIF

Dans cette partie nous décrivons la démarche mise en œuvre dans le cadre de notre recherche d'un modèle prédictif permettant d'anticiper l'ampleur d'un incendie.

Comme nous le verrons, nous serons confrontés à un problème de type classification dans lequel il s'agit de prédire la répartition des incendies de part et d'autre d'un seuil. Ce type d'approche a déjà été utilisé [4,5] pour le Portugal et la Turquie, en mettant en œuvre différents algorithmes de classification tels que les SVM, les réseaux de neurones, les random forest...

Nous traiterons dans le premier chapitre de la préparation des données. Dans les chapitres ultérieurs, nous passerons à l'élaboration du modèle.

5.1 Préparation des données

Comme ceci a été mentionné dans la partie 4, il est préférable de choisir nos données d'entrée parmi les années ultérieures à 1990 si l'on souhaite élaborer un modèle tenant compte au moins implicitement des moyens modernes d'intervention et de prévention. Nous avons finalement retenu les relevés s'étalant du 1^{er} janvier 2002 au 31 décembre 2014. Nous ne sommes pas descendus en deçà de l'année 2002 dans la mesure où les relevés du site wunderground.com présentent de nombreuses lacunes.

Les données fournies par la base Prométhée dont nous disposons dans ce relevé sont, pour chaque feu de forêt : la date et l'heure de l'alerte incendie, la date et l'heure de début et de fin d'intervention, la surface brûlée à l'arrivée des secours, la surface brûlée après intervention, la localisation (département, commune, code INSEE, adresse et carreau DFCI), la cause de l'incendie (involontaire, travaux domestiques ou agricoles, malveillance,...), le type de forêt, la distance des habitations et du réseau routier à l'incendie.

La principale tâche concernant la phase de préparation des données consiste alors à rechercher, pour chacun des incendies ainsi extraits, les conditions météorologiques régnant au moment de l'alerte incendie. Ceci est réalisé en recherchant, pour chaque incendie, l'aéroport le plus proche grâce à leurs coordonnées géographiques, elles-mêmes obtenues par géocodage des communes où ils sont situés. Une fois l'aéroport obtenu, un webscrapping sur le site wunderground est réalisé pour récupérer le bulletin météo à la date de l'alerte.

Pour un incendie donné, nous n'utilisons que les données météo du jour du départ de feu. Pour vérifier la pertinence de ce choix, nous avons calculé la durée en nombre de jours des incendies en fonction de la surface brûlée. Il en ressort que les incendies de forêt dont la durée est comprise entre 1 et 2 jours brûlent en moyenne 10 hectares de forêt. Or, nous utilisons dans la suite un seuil de 5 hectares pour la classification. De ce point de vue, la première journée s'avère déterminante. Le code R associé à ce résultat est présenté en annexe D1.

L'ensemble du code R associé à la préparation des données (géocodage, calcul de distance entre incendie et aéroport) figure dans l'annexe D2. Nous présentons toutefois ici le calcul de la distance entre un incendie et un aéroport. Il fait intervenir leurs coordonnées géographiques ainsi que les caractéristiques géométriques de la Terre. Il existe plusieurs représentations géométriques du globe terrestre, la plus communément utilisée étant la norme WGS84. Elle assimile la Terre à un ellipsoïde de demi-grand axe mesurant 6378,1 kms et de demi-petit axe 6356,7 kms [6]. Les formules donnant les coordonnées (x_A, y_A, z_A) d'un point A sur la surface terrestre dans un repère cartésien géocentré sont les suivantes :

$$\begin{aligned}
x_A &= N_A \cdot \cos(lat_A) \cdot \cos(long_A) \\
y_A &= N_A \cdot \cos(lat_A) \cdot \sin(long_A) \\
z_A &= \left(\frac{a}{b}\right)^2 N_A \cdot \sin(lat_A)
\end{aligned}
\quad \text{Avec : } N_A = \frac{a^2}{\sqrt{a^2 \cos^2(lat_A) + b^2 \sin^2(lat_A)}}$$

a et b désignent, respectivement, la longueur du demi-grand axe et du demi-petit axe de l'ellipsoïde terrestre, tandis que lat_A et $long_A$ désignent la latitude et la longitude du point A. L'aéroport dont la distance euclidienne avec l'incendie est la plus faible lui est alors associé.

Un filtrage des valeurs indéterminées a ensuite été réalisé. Ceci concerne en premier lieu les variables météorologiques, certains historiques étant incomplets. Ces lacunes touchent environ 20% des 24200 incendies répertoriés du 1^{er} janvier 2002 au 31 décembre 2014.

Nous avons ensuite choisi de nous limiter à la période estivale et de ne pas prendre en compte les feux de forêt dont la surface brûlée finale est inférieure à 1 hectare. En effet, les petits feux sont en général accidentels et risquent de fausser les statistiques, leur propagation étant vite maîtrisée, indépendamment des conditions météorologiques, du fait, généralement de la présence humaine lors de l'éclosion du feu.

5.2 Elaboration d'un modèle prédictif à partir des données météo

Nous avons choisi de mettre en œuvre un modèle prédictif de type classification. Comme dans [4,5], le modèle consistera à prédire en fonction des données météorologiques la répartition des incendies de part et d'autre d'un seuil de surface brûlée de 5 hectares.

Nous présentons sur la figure 18 la répartition des incendies de part et d'autre de ce seuil en fonction des données météorologiques. Les échantillons dont la surface brûlée est inférieure à 5 hectares correspondent aux points rouges et les incendies ayant brûlé plus de 5 hectares aux points verts. Vmoy, Tmoy et Hmoy correspondent respectivement à la vitesse du vent moyenne, la température moyenne et le taux d'humidité moyen.

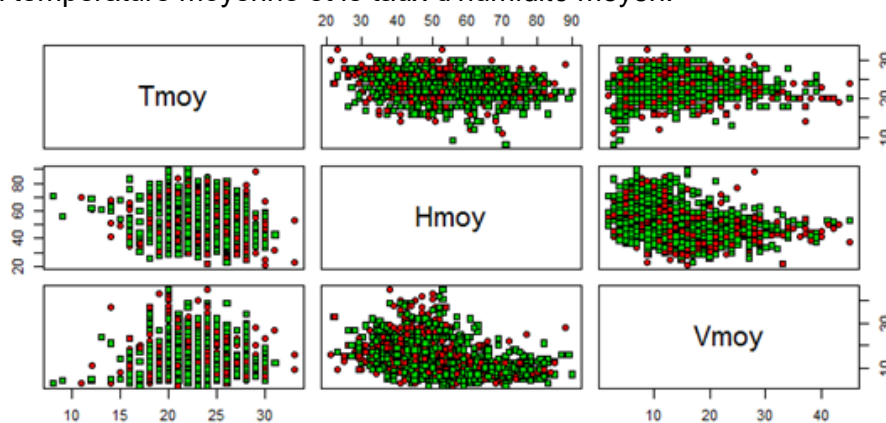


Figure 18 : répartition des incendies en fonction des variables météorologiques.

Nous constatons que les échantillons sont très difficilement séparables. Un modèle prédictif fiable à partir de ces seules données s'annonce donc peu plausible.

Pour confirmer cette intuition, nous avons mis en œuvre un modèle basé sur les SVM. Nous utilisons un noyau gaussien du fait du caractère non linéairement séparable des données [7,8]. Le modèle est entraîné sur les années 2002 à 2010 et sa validation est réalisée sur les années 2011 à 2014. Afin d'optimiser l'algorithme, nous faisons varier entre 0 et 50 le paramètre γ du noyau et le paramètre C contraignant les erreurs de classement puis gardons le couple (γ, C)

donnant le taux d'erreur le plus faible sur les données d'entraînement en cross-validation. La matrice de confusion obtenue en appliquant le modèle résultant aux données de validation est la suivante (Y_{hat} désigne la valeur prédite et Y_{pred} les données réelles) :

	Yhat	
Y_test	-1	1
-1	232	59
1	127	45

Nous constatons un taux de faux négatifs très important (environ 70%). Notre modèle n'est donc pas satisfaisant eu égard aux enjeux du problème dans la mesure où l'intérêt de la prédiction des incendies de forêt réside précisément dans le fait de pouvoir mieux anticiper les grands incendies. L'ensemble du code R associé à ce chapitre est décrit dans l'annexe D3.

Puisque les données météorologiques sont insuffisantes pour obtenir un modèle fiable, nous leur adjoignons maintenant des données supplémentaires.

5.3 Prise en compte des données environnementales

La prise en compte des données environnementales est indispensable si l'on souhaite traiter de manière exhaustive l'ensemble des variables qui préexistent à l'éclosion d'un feu. Ces variables se découpent globalement en deux catégories :

- L'environnement naturel : la végétation (garrigue, conifères,...).
- L'environnement humain : la population et la densité de population.

Nous mettons en évidence leur influence dans les deux paragraphes suivants.

5.3.1 Influence de la végétation

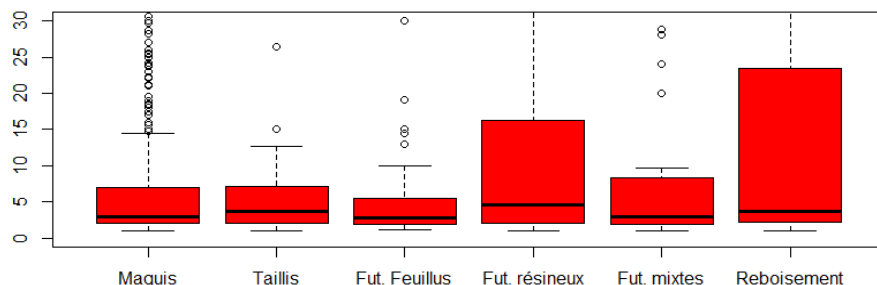
Concernant les informations liées à la végétation, la base Prométhée répartit cette donnée selon les six catégories suivantes : garrigue/maquis, taillis, futaies de résineux, futaies de feuillus, futaies mélangées et zones de reboisement.

Le résultat de l'analyse de variance (ANOVA) de la surface brûlée sur ce facteur est le suivant :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feuxRandom_sav\$TYPE_FORET	1	6.919e+12	6.919e+12	7.732	0.00553 **
Residuals	965	8.636e+14	8.949e+11		

La végétation joue donc un rôle très significatif dans la répartition de la surface moyenne, ce qui se confirme par une très forte disparité de celles-ci pour chacun de ces écosystèmes : 14 ha pour le maquis, 7 ha pour les taillis, 43 ha pour les futaies de résineux, 32 ha pour les futaies de feuillus, 38 ha pour les futaies mixtes et 20 ha pour les zones de reboisement.

Les box-plot suivantes illustrent la répartition des incendies selon ces 6 écosystèmes :



5.3.2 Influence des variables démographiques

En ce qui concerne l'environnement humain, nous pouvons présupposer que les incendies se développent d'autant plus facilement qu'ils surviennent dans une commune rurale présentant une faible population. En effet, plus la présence humaine est rare, moins vite est repéré le feu. La superficie de la commune peut également intervenir dans la mesure où une commune

étendue sera susceptible, pour un nombre d'habitants identique, de présenter plus de massifs forestiers isolés qu'une commune moins étendue. Ces informations (taux de population et superficie) sont disponibles en open data sur le site officiel de l'INSEE.

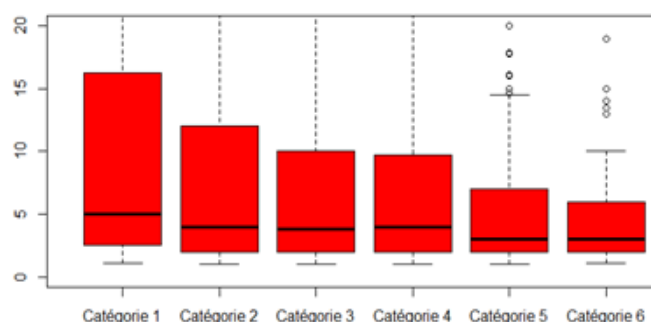
L'influence du taux de population sur la propagation des incendies est mise en évidence ci-après. Nous avons regroupé les communes selon les six catégories suivantes :

- Catégorie 1 : entre 0 et 100 habitants.
- Catégorie 2 : entre 100 et 500 habitants.
- Catégorie 3 : entre 500 et 1500 habitants.
- Catégorie 4 : entre 1500 et 3000 habitants.
- Catégorie 5 : entre 3000 et 6000 habitants.
- Catégorie 6 : plus de 6000 habitants.

L'analyse de variance de la surface brûlée par rapport à ces six catégories aboutit à :

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
feuxRandom_sav\$rural	1	5.881e+13	5.881e+13	8.824	0.00302 **
Residuals	1484	9.891e+15	6.665e+12		

L'influence de la démographie sur la disparité de la surface moyenne est donc significative, ce qui se confirme par les valeurs moyennes suivantes : 55 ha (cat. 1), 22 ha (cat. 2), 19 ha (cat. 3), 12 ha (cat. 4), 10 ha (cat. 5) et 8 ha (cat. 6). La surface brûlée moyenne décroît donc globalement avec le nombre d'habitants, comme l'illustrent les box-plots suivantes :



5.3.3 Résultats

Comme précédemment, nous avons mis en œuvre un modèle en nous limitant de nouveau à la période estivale, mais en considérant cette fois les variables suivantes : la météo, la population et la superficie des communes ainsi que la végétation et le département (en ce qui concerne les disparités des statistiques d'incendies d'un département à l'autre, cf. partie 1 et en particulier la figure 6).

Le mélange de variables qualitatives (démographie, végétation,...) et de variables quantitatives (météo,...) rendent les random forest particulièrement adaptés au problème. Un calcul a été effectué avec 5000 arbres. Malheureusement, nous observons toujours un taux de faux négatifs (60%) rédhibitoire, comme le montre la matrice de confusion résultante :

	-1	1	class.error
-1	698	219	0.2388222
1	329	216	0.6036697

Un modèle prédictif basé sur les SVM a également été mis en œuvre, mais fournit des taux d'erreur similaires. L'ensemble du code associé à ce chapitre est présenté en annexe D4, SVM y compris.

Nous ne sommes toujours pas parvenus à établir un modèle prédictif. Toutefois, parmi les variables disponibles au sein de la base Prométhée mais non encore utilisées, figurent les variables caractérisant le déroulement de l'intervention. Le chapitre suivant a pour objectif d'en extraire un modèle prédictif.

5.4 Prise en compte des données d'intervention

Parmi les variables disponibles au sein de la base Prométhée et caractérisant les interventions, figurent notamment la date et l'heure de l'arrivée des secours ainsi que la date et l'heure de la fin de l'intervention. Couplées avec la date et l'heure de l'alerte incendie, nous pouvons alors calculer le délai d'arrivée des secours ainsi que la durée de l'intervention et de l'incendie. La surface déjà brûlée à l'arrivée des secours est également disponible.

Nous mettons ci-après en évidence la présence ou non de corrélation entre ces variables avant de passer à l'élaboration d'un modèle prédictif.

5.4.1 Corrélations entre les variables liées à l'intervention

Puisque de nombreux incendies sont de petite taille, nous remplaçons la surface brûlée par son logarithme [4,5] afin de lisser cette variable.

En considérant également le logarithme de la surface à l'arrivée des secours, nous obtenons les coefficients de corrélation linéaire suivants :

	Surface Finale	Surface initiale	Délai	Durée
Surface Finale	1.000	0.890	0.098	0.636
Surface initiale	0.890	1.000	0.114	0.534
Délai	0.098	0.114	1.000	0.220
Durée	0.636	0.534	0.220	1.000

Nous constatons une corrélation très importante (environ 0,9) entre le logarithme de la surface brûlée après intervention et le logarithme de la surface brûlée à l'arrivée des secours. Ensuite, on remarque que cette dernière est très peu corrélée avec le délai d'intervention ; l'explication est simple : dans la plupart des cas, le délai d'arrivée des forces d'intervention est inférieur à 50 minutes, comme le montre l'histogramme de la figure 19 ; le délai d'intervention est donc en général suffisamment court pour peu influencer sur l'extension du feu. Enfin, logiquement, la durée de l'incendie est assez corrélée avec la surface brûlée finale (coefficient valant 0.63).

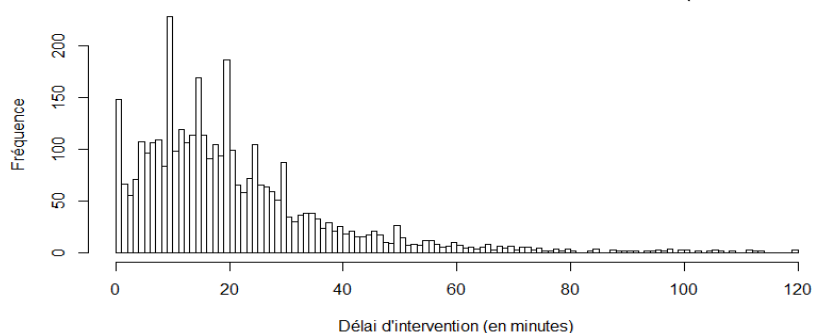


Figure 19 : répartition des incendies en fonction du délai d'intervention.

Une régression linéaire entre le logarithme de la surface brûlée après intervention et le logarithme de la surface brûlée à l'arrivée des secours donne les coefficients suivants :

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.04883 0.03416 30.70 <2e-16 ***
log(feuxRandom_sav$SURFACE_DEB/10000) 0.98105 0.01091 89.89 <2e-16 ***
---
signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.103 on 2012 degrees of freedom
```

La pente de la droite est donc de l'ordre de 1 et l'ordonnée à l'origine de l'ordre de 1,05. Nous avons donc, en moyenne, la relation suivante : $\ln S_{finale} = \ln S_{initiale} + 1,05$.

Soit finalement, en moyenne : $S_{finale} \approx 3.S_{initiale}$

Soulignons bien qu'il s'agit là d'un coefficient moyen. En effet, l'écart-type des valeurs est très important ($\approx 1,1$, cf. ci-dessus) et s'avère être proche de l'ordonnée à l'origine, donc du coefficient lui-même.

La figure 20 récapitule ces résultats. En vert est tracé l'intervalle de confiance de la régression.

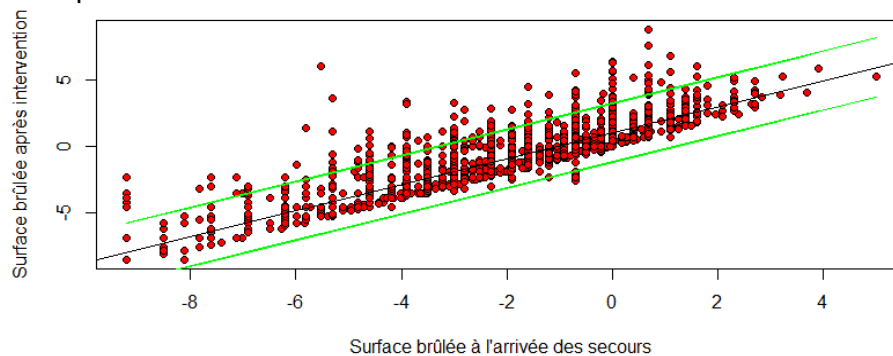


Fig. 20 : régression linéaire entre surface brûlée avant l'arrivée des secours et après l'intervention.

Le code associé à ce paragraphe est présenté en annexe D5.

5.4.2 Elaboration d'un modèle prédictif

Nous nous appuyons maintenant sur les résultats précédents en vue d'établir un nouveau modèle prédictif. Parmi les données concernant l'intervention, nous nous focalisons sur la surface brûlée à l'arrivée des secours.

Nous avons relevé que la surface brûlée à la fin de l'intervention est en moyenne de l'ordre du triple de la surface initiale. Comme nous l'avons fait remarquer, il s'agit là d'une moyenne et l'écart type est très important. Toutefois, nous pouvons partir de cela pour reformuler notre problématique : l'objectif pourrait consister à prédire si la surface brûlée risque de dépasser ou non le triple de la surface brûlée initiale en ajoutant à cette dernière variable les mêmes données que dans le chapitre précédent (météo, végétation,...) avec le jour de la semaine en plus. Cette approche aiderait alors à la décision de prévoir des renforts ou non.

Nous utilisons tout d'abord l'algorithme random forest car il permet d'évaluer l'importance de chaque variable concernant la prédiction. Nous avons choisi un nombre de 5000 arbres. Comme le graphe suivant le montre, cette quantité est suffisante car le taux d'erreur se stabilise à partir de 3000 arbres :

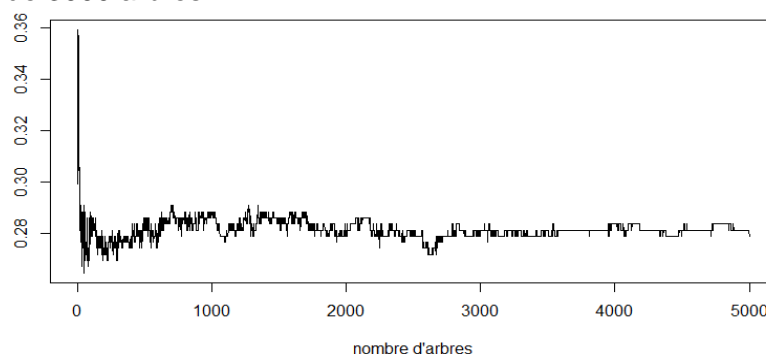


Figure 21 : évolution du taux d'erreur en fonction du nombre d'arbres.

La matrice de confusion qui en résulte est la suivante :

	-1	1	class.error
-1	279	150	0.3496503
1	135	403	0.2509294

Le taux d'erreur total est de 30 % et le taux de vrais positifs de 25 %. Ces quantités se rapprochent des taux d'erreur mentionnés dans [4,5]. De plus, nous pouvons tout à fait imaginer que ces résultats sont susceptibles de s'améliorer considérablement en disposant des historiques des stations Météo-France qui maillent évidemment beaucoup mieux le territoire que les aéroports et sont donc plus précises. En effet, les données météo restent le facteur le plus déterminant après la surface brûlée initiale, comme le montre le diagramme d'importance des variables de la figure 22. Plus précisément, parmi les trois variables météo, c'est le taux d'humidité qui se détache le plus. Viennent ensuite le département et le jour de la semaine. Enfin, nous constatons que la végétation est la variable la moins influente.

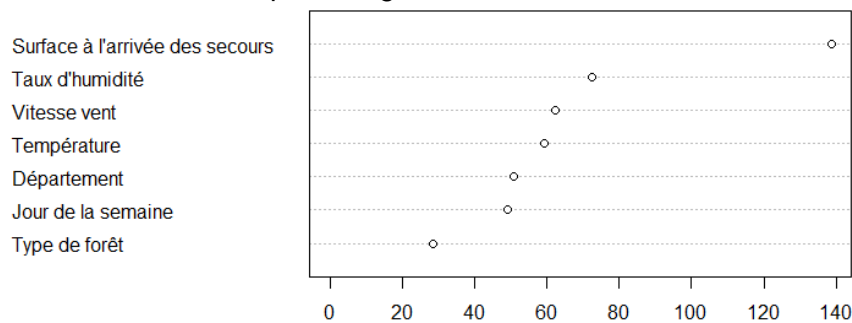


Figure 22 : classement des variables explicatives.

Avec les SVM (entraînés sur les années 2002 à 2010 et validés sur les années 2010 à 2014), nous obtenons la matrice de confusion suivante :

	Yhat	
Y_test	-1	1
-1	79	23
1	40	120

Le résultat est encore meilleur que celui obtenu par les random forest dans la mesure où le taux de vrais positifs passe à 80 %.

Pour souligner l'importance des données météorologiques dans le modèle, nous aboutissons à la matrice de confusion suivante en leur absence :

	-1	1	class.error
-1	344	85	0.1981352
1	238	300	0.4423792

Nous observons alors un taux de faux négatifs de 44 %.

5.5 Bilan

Cette partie avait pour objectif de présenter un modèle prédictif fiable permettant de mieux anticiper l'ampleur d'un incendie.

La première approche a consisté dans un premier temps à évaluer le risque de dépassement d'une superficie seuil soit à partir des seules données météorologiques soit en leur ajoutant des données supplémentaires qui concernent principalement l'environnement du feu. Malheureusement, nous ne sommes pas parvenus à établir un modèle fiable.

Il a donc été décidé de faire intervenir dans le modèle des variables liées à l'intervention des secours elle-même. Ces données sont disponibles dans la base Prométhée et comprennent notamment la surface brûlée à l'arrivée des secours. La problématique a alors été légèrement modifiée : il s'agit cette fois de prédire dans quelles conditions la surface d'un feu de forêt risque d'au moins tripler après l'arrivée des secours. Nous avons abouti à un modèle prédictif fiable que ce soit avec les random forest ou les SVM dans lesquels la météo joue bien un rôle déterminant.

6 CONCLUSION

Nous avons étudié dans ce document la prédictibilité des feux de forêt dans le quart sud-est de la France en fonction des conditions météorologiques.

Dans la première partie, nous avons dégagé les principaux enjeux économiques et environnementaux de cette problématique. A travers le réchauffement climatique, nous y avons notamment mis en évidence l'extension des zones à risque sur l'ensemble du territoire français et, par conséquent, le coût économique croissant des feux de forêt durant les décennies à venir. Ceci nous a donc confortés dans l'idée de l'importance de la modélisation des feux de forêt à partir des conditions météorologiques.

Dans la seconde partie, ont été brièvement présentées les deux sources de données à partir desquelles nous avons extrait l'ensemble des données relatives aux feux de forêt et aux données météorologiques pour la construction de notre modèle. Ainsi, le site web promethee.com recense l'ensemble des caractéristiques des incendies (date et heure de l'alerte, du début et de la fin de l'intervention, cause, surface de forêt brûlée, commune, géolocalisation,...) tandis que le site web wunderground.com nous fournit les données météorologiques concomitantes à chaque incendie.

Dans la troisième partie, nous avons mis en œuvre la base de données orientée document [mongodb](https://www.mongodb.com/) en vue d'extraire à partir des relevés issus de ces deux sites un échantillon pertinent d'incendies avec les données météorologiques associées.

Les échantillons ainsi obtenus ont ensuite été exploités dans la quatrième partie dans le but de faire ressortir quelques grandes tendances concernant les feux de forêt. Nous y avons notamment comparé les données concernant les incendies survenus en 2003 et en 2014 afin de mettre en évidence l'influence de la température sur le développement des incendies.

La cinquième et dernière partie était consacrée à la recherche d'un modèle prédictif. Nous y avons tout d'abord présenté la méthodologie par laquelle nous avons associé à chaque incendie les conditions météorologiques concomitantes grâce, notamment, au webscrapping. Ensuite, nous avons tenté de dégager, à partir de ces seules conditions un modèle prédictif. Celles-ci se sont alors révélées insuffisantes pour établir un modèle fiable. D'autres données environnementales leur ont donc été ajoutées, en vain. Finalement, les variables liées à l'intervention des secours se sont avérées décisives. Un modèle prenant en compte la surface brûlée à l'arrivée des secours a ainsi été élaboré. Deux approches ont été mises en œuvre : les random forest et les SVM. Les SVM donnent le meilleur résultat.

Contrairement aux études déjà réalisées au Portugal et en Turquie, l'élaboration d'un modèle similaire a été plus compliquée pour la France et a nécessité de prendre en compte des variables supplémentaires. Cette difficulté a des raisons propres à la France qui sont notamment liées aux nombreux investissements réalisés pour la prévention des feux de forêt (tours de guet, avions de reconnaissance,...) ainsi qu'aux mesures répressives mises en place en période estivale (fermeture des sentiers de randonnées durant les périodes à risque,...). Les grands feux de forêt sont donc condamnés à être en France de plus en plus aléatoires.

Enfin, notons que le taux d'erreur du modèle final (de l'ordre de 20 %) peut être largement amélioré si l'on utilise les historiques des stations Météo-France au lieu des historiques des aéroports parfois éloignés d'une centaine de kilomètres des feux et donc beaucoup moins précis géographiquement.

BIBLIOGRAPHIE

[1] : Conservatoire de la forêt méditerranéenne, « Emploi du feu et débroussaillage en Ardèche ». Année : 2006. Disponible à l'adresse suivante :

www.debroussaillage-foret.ardeche.agriculture.gouv.fr/IMG/pdf/Emploi_Feu_Debrouss_Ardeche_cle0d9eea.pdf

[2] : Commissariat général au développement durable, « Le risque de feux de forêt en France ». Août 2011. Disponible à l'adresse suivante :

www.developpement-durable.gouv.fr/IMG/pdf/ED45-2.pdf

[3] : Jean-Luc Dupuis, Institut National de la Recherche Agronomique, « Propagation et impacts des feux de végétation : enjeux et modélisation ». Disponible à l'adresse suivante :

http://www.coriolis.polytechnique.fr/Confs/Dupuy_conf.pdf

[4] : Paulo Cortez, Anibal Morais, Department of information systems/R&D algorithmi centre, University of Minho, Portugal, « A Data Mining Approach to Predict Forest Fires using Meteorological Data ». Conférence EPIA. Année : 2007.

[5] : A. Murat Ozbayoglu, Recep Bozer, TOBB University of Economics and Technology, Department of Computer Engineering, Ankara, 06560, Turkey, « Estimation of the burned area in forest fires using computational intelligence techniques ». Elsevier, Procedia Computer Science, volume12, pp 282–287. Année : 2009.

[6] : Institut géographique national : « Transformations géodésiques en France métropolitaine. ». Année : 2013. Disponible à l'adresse suivante :

<http://geodesie.ign.fr/contenu/fichiers/documentation/pedagogiques/TransformationsCoordonneesGeodesiques.pdf>

[7] : Trevor Hastie, Robert Tibshirani et Jerome Friedman, « The elements of statistical learning : data mining, inference and prediction (second edition) ». Editions Springer. Année : 2009. Disponible à l'adresse suivante :

http://statweb.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf

[8] : Gilbert Saporta, « Probabilités, analyse des données et statistique (seconde édition) ». Editions Technip. Année 2006.

ANNEXES

A. CODE R ASSOCIE A LA PARTIE 1

A1. Partitionnement des différents départements de la zone Prométhée

```
library(dplyr) # Manipulation simple et rapide des données dans les dataframes notamment
library(lubridate) # Manipulation des dates
library(stringr) # Manipulation des chaînes de caractères
library(reshape2) # Agrégation et restructuration de données
library(factoextra) # Librairie d'analyse de données contenant notamment l'analyse en composantes principales
```

```
feux = read.csv("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/feux_aeroports_avec
Meteo_lastt_avecNA_causes.csv", header=TRUE, sep=";")
```

```
# Parsing de la date au format DATE de R
```

```
feux$DATE_HEURE = parse_date_time(feux$DATE, '%d/%m/%y %h:%m')
```

```
DATE_JOUR <- NULL
```

```
# On enlève l'heure de la date
```

```
for (i in 1:length(feux$ANNEE))
```

```
{
```

```
  DATE_JOUR[i] = as.character(str_split(feux$DATE[i], " ")[[1]][1], orders="dmy")
```

```
}
```

```
feux$DATE_JOUR = DATE_JOUR
```

```
# Extraction du mois à partir de la date
```

```
month = month(feux$DATE)
```

```
feux$month = month
```

```
# Extraction du code postal et du département associé aux feux
```

```
feux$CP = str_c("0", feux$CP)
```

```
feux$CP = str_sub(feux$CP, -5)
```

```
feux$dep = NULL
```

```
code_dep = c("83", "2A", "26", "30", "84", "04", "06", "11", "2B", "13", "07", "48", "05", "66", "34")
```

```
no_dep = c(1:15)
```

```
for (i in 1:length(feux$CP))
```

```
{
```

```
  feux$dep[i] = no_dep[which(code_dep == str_sub(feux$CP[i], 1, 2))]
```

```
}
```

```
feux_sav = feux
```

On agrège le nombre d'incendies sur les départements et les 12 mois de l'année

```
feuxMois = aggregate(feux_sav$SURFACE/10000.0, by=feux_sav[c("month","dep")], FUN="length")
```

On convertit le data frame obtenu en tableau avec en ligne les départements et en colonnes les 12 mois de l'année

```
tabMoisDep = acast(feuxMois,dep~month)
```

Cas où aucun incendie n'a eu lieu pour un département donné durant un mois donné

```
tabMoisDep[is.na(tabMoisDep)] = 0
```

On somme le nombre d'incendies pour chaque département

```
somme = apply(tabMoisDep,1,sum)
```

On normalise chaque ligne par le nombre total d'incendie sur cette ligne pour obtenir des pourcentages par mois

```
for (i in 1:15){  
  for (j in 1:12){  
    tabMoisDep[i,j] = tabMoisDep[i,j]*100.0 / somme[i]  
  }  
}
```

Nommage des lignes et colonnes

```
colnames(tabMoisDep)=("Janvier","Février","Mars","Avril","Mai","Juin","Juillet","Août","Septembre","Octobre","Novembre","Décembre")
```

```
rownames(tabMoisDep) = code_dep
```

Calcul de l'analyse en composantes principales sur le tableau précédent

```
res.pca = prcomp(tabMoisDep, scale. = TRUE)
```

```
summary(res.pca)
```

Affichage de la variance expliquée par chacun des 12 axes factoriels obtenus

Les 2 premiers axes factoriels expliquent 64% de la variance totale

```
plot(res.pca, main = "explained variance")
```

Clustering ascendant hiérarchique sur le tableau précédent

Distance : norme euclidienne standard

Ultramétrie : distance de Ward

```
res.hclust = hclust(dist(tabMoisDep, method = "euclidean")^2, method = "ward.D")
```

Affichage du dendrogramme

```
plot(res.hclust)
```

```
rect.hclust(res.hclust,2)
```

Affichage du bi-plot associé aux deux axes factoriels de l'ACP

**# Les départements sont affichés avec 2 couleurs différentes qui correspondent aux 2 groupes
résultant de la coupure en 2 groupes du dendrogramme**

```
fviz_pca_biplot(res.pca, col.ind = cutree(res.hclust,2), repel = TRUE)
```

On constitue finalement 3 grands groupes de département résultant de l'ACP et du clustering hiérarchique : Corse, départements littoraux et départements de l'intérieur

```
vec_med = which(feux$dep %in% c(1,4,5,8,10,14,15))
feux_littoraux <- feux_sav[-vec_med,]
vec_corse = which(feux$dep %in% c(2,9))
feux_corse <- feux_sav[-vec_corse,]
vec_montagne = which(feux$dep %in% c(3,6,7,11,12,13))
feux_montagne <- feux_sav[-vec_montagne,]
```

On compte, pour chaque groupe de départements, le nombre total d'incendies pour chaque mois

```
feuxMoisLitt = aggregate(feux_littoraux$aeroport, by=feux_littoraux[c("month")], FUN="length")
feuxMoisMontagne=aggregate(feux_montagne$aeroport,by=feux_montagne[c("month")],
FUN="length »)
feuxMoisCorse = aggregate(feux_corse$aeroport, by=feux_corse[c("month")], FUN="length")
```

Affichage sur le même graphe du nombre d'incendies par mois pour chaque département littoral de la zone Prométhée

```
feuxVar = feuxMois[which(feuxMois$dep == 1),]
plot(feuxVar$month, feuxVar$x, pch = 21, col = "blue", xlab = "Mois", ylab = "Nombre d'incendies",
xlim=c(1,12), type="l", main="Littoral", ylim=c(0,230))
par(new=TRUE)
feuxGard = feuxMois[which(feuxMois$dep == 4),]
plot(feuxGard$month, feuxGard$x, pch = 21, col = "green", xlab = "Mois", ylab = "Nombre d'incendies",
xlim=c(1,12), type="l", main="Littoral",ylim=c(0,230))
feuxVaucluse = feuxMois[which(feuxMois$dep == 5),]
par(new=TRUE)
plot(feuxVaucluse$month, feuxVaucluse$x, pch = 21, col = "cyan", xlab = "Mois", ylab = "Nombre
d'incendies", xlim=c(1,12), type="l", main="Littoral",ylim=c(0,230))
feuxAude = feuxMois[which(feuxMois$dep == 8),]
par(new=TRUE)
plot(feuxAude$month, feuxAude$x, pch = 21, col = "red", xlab = "Mois", ylab = "Nombre d'incendies",
xlim=c(1,12), type="l", main="Littoral",ylim=c(0,230))
feuxBouches = feuxMois[which(feuxMois$dep == 10),]
par(new=TRUE)
plot(feuxBouches$month, feuxBouches$x, pch = 21, col = "purple", xlab = "Mois", ylab = "Nombre
d'incendies", xlim=c(1,12), type="l", main="Littoral",ylim=c(0,230))
feuxPyr = feuxMois[which(feuxMois$dep == 14),]
```



```

par(new=TRUE)

plot(feuxPyr$month, feuxPyr$x, pch = 21, col = "yellow", xlab = "Mois", ylab = "Nombre d'incendies",
xlim=c(1,12), type="l", main="Littoral",ylim=c(0,230))

feuxHerault = feuxMois[which(feuxMois$dep == 15),]

par(new=TRUE)

plot(feuxHerault$month, feuxHerault$x, pch = 21, col = "black", xlab = "Mois", ylab = "Nombre
d'incendies", xlim=c(1,12), type="l", main="Littoral",ylim=c(0,230))

legend(x=9.3, y=240, legend=c("Bouches du Rhône", "Hérault", "Aude", "Var", "Gard", "Pyrénées
Orientales", "Vaucluse"), col=c("purple", "black", "red", "blue", "green", "yellow", "cyan"), lty = 1,cex=0.8)

# Affichage sur le même graphe du nombre d'incendies par mois pour chaque département de
l'intérieur de la zone Prométhée

feuxDrome = feuxMois[which(feuxMois$dep == 3),]

plot(feuxDrome$month, feuxDrome$x, pch = 21, col = "cyan", xlab = "Mois", ylab = "Nombre
d'incendies", xlim=c(1,12), type="l", main="Montagne", ylim=c(0,100))

par(new=TRUE)

feuxHP = feuxMois[which(feuxMois$dep == 6),]

plot(feuxHP$month, feuxHP$x, pch = 21, col = "green", xlab = "Mois", ylab = "Nombre d'incendies",
xlim=c(1,12), type="l", main="Montagne",ylim=c(0,100))

feuxAM = feuxMois[which(feuxMois$dep == 7),]

par(new=TRUE)

plot(feuxAM$month, feuxAM$x, pch = 21, col = "blue", xlab = "Mois", ylab = "Nombre d'incendies",
xlim=c(1,12), type="l", main="Montagne",ylim=c(0,100))

feuxArdeche = feuxMois[which(feuxMois$dep == 11),]

par(new=TRUE)

plot(feuxArdeche$month, feuxArdeche$x, pch = 21, col = "red", xlab = "Mois", ylab = "Nombre
d'incendies", xlim=c(1,12), type="l", main="Montagne",ylim=c(0,100))

feuxLozere = feuxMois[which(feuxMois$dep == 12),]

par(new=TRUE)

plot(feuxLozere$month, feuxLozere$x, pch = 12, col = "purple", xlab = "Mois", ylab = "Nombre
d'incendies", xlim=c(1,12), type="l", main="Montagne",ylim=c(0,100))

feuxHA = feuxMois[which(feuxMois$dep == 13),]

par(new=TRUE)

plot(feuxHA$month, feuxHA$x, pch = 13, col = "yellow", xlab = "Mois", ylab = "Nombre d'incendies",
xlim=c(1,12), type="l", main="Montagne",ylim=c(0,100))

legend(x=9, y=100, legend=c("Lozère", "Ardèche", "Alpes Maritimes", "Alpes Haute Provence", "Hautes
Alpes", "Drôme"), col=c("purple", "red", "blue", "green", "yellow", "cyan"), lty = 1,cex=0.8)

# Affichage sur le même graphe du nombre d'incendies survenu en Corse pour chaque mois

feuxCorseA = feuxMois[which(feuxMois$dep == 2),]

feuxCorseB = feuxMois[which(feuxMois$dep == 9),]

feuxCorseA$x = feuxCorseA$x + feuxCorseB$x

```

```
plot(feuxCorseA$month, feuxCorseA$x, pch = 21, col = "blue", xlab = "Mois", ylab = "Nombre d'incendies", xlim=c(1,12), type="l", main="Corse",ylim=c(0,450))
```

A2. Analyse de la surface brûlée par département.

```
library(stringr) # Manipulation des chaînes de caractères
```

```
library(ggplot2) # Outils de visualisation graphiques
```

```
# Lecture des échantillons d'incendies de la zone Prométhée de 2010 à 2014
```

```
feux = read.csv("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/feux_aeroports_avecMeteo_lastt_avecNA_causes.csv", header=TRUE, sep=";")
```

```
# Codes INSEE des départements de la zone Prométhée
```

```
code_dep <- c("05","04","06","83","84","13","2A","2B","30","26","07","48","34","11","66")
```

```
# Noms des départements associés à chacun des codes INSEE du vecteur code_dep
```

```
Dep <- c("Hautes Alpes","Haute Provence","Alpes Maritimes","Var","Vaucluse","Bouches du Rhône",  
"Corse du Sud", "Haute Corse", "Gard", "Drôme", "Ardèche", "Lozère", "Hérault", "Aude", "Pyrénées  
Orientales")
```

```
feux$dep <- NULL
```

```
# On ajoute le 0 au début pour les départements dont le code est INSEE est < 10
```

```
feux$CP = str_c("0",feux$CP)
```

```
feux$CP = str_sub(feux$CP,-5)
```

```
# On extrait le numéro du département de l'incendie à partir de son code postal
```

```
for (i in 1:length(feux$CP))
```

```
{
```

```
  feux$dep[i] = Dep[which(code_dep == str_sub(feux$CP[i],1,2))]
```

```
}
```

```
head(feux$dep,100)
```

```
# On calcule la surface moyenne (en hectares) de chaque incendie par département
```

```
statDep = aggregate(feux$SURFACE/10000.0, by=list(feux$dep), FUN="mean")
```

```
statDep =statDep[order(statDep[,2]), ]
```

```
# Tracé de l'histogramme représentant la surface brûlée par incendie pour chaque département
```

```
ggplot(data = statDep,aes(x=Group.1, y=x)) + geom_bar(stat="identity", color="black", fill="blue") +  
coord_flip() + ylab("Surface brûlée moyenne (en hectares)")
```

B. CODE R ASSOCIE A L’AFFICHAGE DE LA CARTE DES AEROPORTS (PARTIE 2)

Ce code est à l’origine des cartes des figures 10a et 10b. Elles représentent l’ensemble des aéroports du sud-est de la France dont l’historique des relevés météorologiques est disponible sur le site www.wunderground.com.

```
library(rvest) # librairie contenant des outils de webscrapping

library(ggmap) # librairie contenant des fonctions de visualisation de données spatiales et de
calculs de positions et d’itinéraires

library(RgoogleMaps) # affichage de fonds de carte issus de googlemap

# URL où se trouve la liste complète des aéroports français dont on veut extraire les données
météorologiques

list.url<- http://www.wunderground.com/history/index.html?error=AMBIGUOUS&query=France&day
=17&month=8&year=2015&MR=1"

# Liste des balises contenant les liens vers chacun des aéroports

node <- "#history-station-search-row li a"

# Capture des liens vers les pages des aéroports

airports.list <- read_html(list.url) %>% html_nodes(node) %>% html_attr("href")

# Capture du texte associé au lien (nom de la ville où se situe l’aéroport)

airports.names <- read_html(list.url) %>% html_nodes(node) %>% html_text

# On ne retient dans cette liste que les liens vers les aéroports de la zone Prométhée ou proches
de la zone Prométhée et possédant un historique météorologique. Chaque aéroport est en fait
défini par la ville à laquelle il est rattaché

airports_hist <- airports.names[c(4,5,14,33,35,45,49,70,73,75,78,80,95,106,109,121,128,129,132,
142,154,161,167,168,170,176)]

# On remplace l’aéroport d’Embrun (inexistant après vérification) par celui de Levaldigi

airports_hist[8] = "Levaldigi, Italy"

# On géocode chaque aéroport à partir de la ville associée à l’aéroport.

latlong <- geocode(airports_hist)

# Affichage du fond de carte issu de googlemap associé à la la figure 5a

France <- GetMap(center=c(44, 4.5), zoom=7,matype="terrain")

# Affichage sur ce fond de carte de chaque aéroport sous forme de points rouge (ou bleu pour
Levaldigi)

PlotOnStaticMap(France, lat = as.numeric(latlong$latitude), lon = as.numeric(latlong$longitude),cex =
2, pch = 19, col = "red", add = FALSE)

PlotOnStaticMap(France, lat = as.numeric(latlong_Levaldigi$latitude), lon =
as.numeric(latlong_Levaldigi$longitude),cex = 2, pch = 19, col = "blue", add = TRUE)

# On importe le fichier de forme (format de fichier pour les systèmes d’information
géographiques) définissant les frontières des départements français.
```

```
shp<-importShapefile("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/dep/departements-20140306-100m.shp",readDBF=FALSE)
```

Affichage sur le fond de carte des frontières sous forme de lignes noires

```
PlotPolysOnStaticMap(France,shp,lwd=1., col=rgb(0,0,0,0), add = T)
```

Affichage des aéroports présents sur le territoire Corse (figure 5b)

```
France <- GetMap(center=c(42.5, 6.9), zoom=7,maptype="terrain")
```

```
PlotOnStaticMap (France, lat = as.numeric(latlong$latitude), lon = as.numeric(latlong$longitude),cex = 2, pch = 19, col = "red", add = FALSE)
```

C. PRESENTATION DU CODE R DE LA PARTIE 4

C1. Analyse des feux de forêt depuis 1973

```
library(stringr) # Librairie de manipulation de chaînes de caractères
library(lubridate) # Librairie de manipulation de dates
library(beeswarm) # Affiche tous les échantillons sur un graphique sous forme de points distincts.
# Ouverture du fichier répertoriant l'ensemble des incendies de 1973 à 2014
feux_1973_2014<-read.delim("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/
feux_1973_2014.csv")
# On compte le nombre de feux de forêt pour chaque année
res = as.data.frame(table(feux_1973_2014$Annee))
# On affiche l'histogramme associé (nombre de feux de forêt en ordonnée et année en abscisse)
barplot( res$Freq, names.arg = res$Var1, col='blue', main="Nombre de feux de forêt annuel")
# On additionne la surface de forêt brûlée totale pour chaque année
res=as.data.frame(aggregate(feux_1973_2014$Surface..m2.~feux_1973_2014$Annee, FUN = "sum"))
# On renomme les deux colonnes du tableau résultant
names(res) <- c("Annee", "Surface")
# On affiche l'histogramme associé (surface brûlée totale en km² en ordonnée et année en abscisse)
barplot(res$Surface/1000000.0,names.arg=res$Annee,col='blue',main='Surface brûlée annuelle (en km²)')
```

C2. Analyse des échantillons de Haute Corse

```
# On ouvre le fichier répertoriant les feux de forêt survenus en 2014 et les conditions
météorologiques associées (cf. partie 3)
feux_corse_2014_meteo<-read.csv("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/
feux_corse_2014_meteo.csv", sep=";")
# Conversion m² -> hectare pour la surface brûlée
feux_corse_2014_meteo$Surface = feux_corse_2014_meteo$Surface/10000.0
# Tracé de l'histogramme représentant la répartition des feux de forêt en fonction de la surface
brûlée. La surface brûlée maximale fut en 2014 de l'ordre de 500 hectares, d'où les 500
échantillons en abscisse. La limite supérieure visible en abscisse est toutefois fixée à 20
hectares, les feux de forêt ayant brûlé plus de 20 hectares de forêt étant très peu nombreux.
hist(feux_corse_2014_meteo$Surface,breaks=500,xlim=c(0,20),col='blue',xlab='Surface (en hectares)'
,ylab='Répartition')
# On sélectionne les incendies de forêt, c'est-à-dire les feux ayant brûlé plus d'un hectare de
forêt
incendies = feux_corse_2014_meteo$Surface[feux_corse_2014_meteo$Surface > 1.0]
```

Affichage de la box-plot associée à la surface brûlée par ces incendies. On fixe une limite visible en ordonnée à 100 hectares pour des raisons de lisibilité des résultats.

```
boxplot(incendies, col = "red", main = "Surface brûlée (en hectares) en 2014",ylim=c(0,100))
```

Affichage des points associés à chacun des incendies par-dessus la box-plot afin de mieux visualiser la répartition des incendies

```
beeswarm(incendies, pch = 21, bg = "black", add = TRUE)
```

Calcul de la température moyenne pour l'ensemble des incendies de forêt.

```
incendies_meteo = filter(feux_corse_2014_meteo, Surface > 1.0)
```

```
Tmoy = mean(incendies_meteo$Tmax)
```

D. PRESENTATION DU CODE R DE LA PARTIE 5

D1. Répartition de la durée des incendies de forêt en fonction de la surface brûlée

```
library(stringr) # Manipulation des chaînes de caractères.
```

```
library(dplyr) # Manipulation simple et rapide des données dans les dataframes notamment
```

```
library(lubridate) # librairie permettant la manipulation des dates
```

```
# Lecture des données fournies par le réseau Prométhée
```

```
feux = read.csv("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/Donnees_SDIS_1996-2015_Duree_intervention.csv", header=TRUE, sep=";")
```

```
# On renomme les deux colonnes associées à la date de début et de fin d'incendie
```

```
names(feux)[4]="DEBUT"
```

```
names(feux)[5]="FIN"
```

```
# Conversion des dates de début et de fin d'incendie au format correspondant aux données fournies
```

```
feux$date_debut = parse_date_time(feux$DEBUT, '%d/%m/%y %h:%M')
```

```
feux$date_fin = parse_date_time(feux$FIN, '%d/%m/%y %h:%M')
```

```
# Calcul de la durée en jours de chaque incendie
```

```
duree <- NULL
```

```
for (i in 1:length(feux$DEBUT))
```

```
{
```

```
  duree[i] = difftime(feux$date_fin[i], feux$date_debut[i], units="days")
```

```
}
```

```
# On arrondit la durée en jour à la valeur entière inférieure
```

```
feux$duree = round(duree)
```

```
head(feux$duree)
```

```
# On calcule puis on affiche la surface moyenne de forêt brûlée pour les incendies durant entre 0 et 1 jour, entre 1 et 2 jours, entre 2 et 3 jours, etc...
```

```
vec1 = which(feux$duree == 0.0)
```

```
vec2 = which(feux$duree == 1.0)
```

```
vec3 = which(feux$duree == 2.0)
```

```
vec4 = which(feux$duree == 3.0)
```

```
vec5 = which(feux$duree == 4.0)
```

```
vec6 = which(feux$duree == 5.0)
```



```

vec7 = which(feux$ duree == 6.0)

mean1 <- mean(feux[vec1,]$Surface.parcourue..m2.)/10000
mean2 <- mean(feux[vec2,]$Surface.parcourue..m2.)/10000
mean3 <- mean(feux[vec3,]$Surface.parcourue..m2.)/10000
mean4 <- mean(feux[vec4,]$Surface.parcourue..m2.)/10000
mean5 <- mean(feux[vec5,]$Surface.parcourue..m2.)/10000
mean6 <- mean(feux[vec6,]$Surface.parcourue..m2.)/10000
mean7 <- mean(feux[vec7,]$Surface.parcourue..m2.)/10000

mean1
mean2
mean3
mean4
mean5

```

D2. Préparation des données à traiter

```

library(stringr) # Manipulation des chaînes de caractères.

library(dplyr)  # Manipulation simple et rapide des données dans les dataframes notamment

library(rvest)  # librairie contenant des outils de webscrapping

library(ggmap)  # librairie contenant des fonctions de visualisation de données spatiales et de
calculs de positions et d'itinéraires

library(NISTunits) # librairie fournissant les valeurs des constantes physiques fondamentales et
permettant aussi de convertir des unités entre elles.

Library(lubridate) # librairie permettant la manipulation des dates

# Fonction qui calcule la distance entre l'aéroport et l'incendie selon la norme WGS84. La formule
est donnée dans le chapitre 5.1

# LatAero et longAero désignent la latitude et la longitude de l'aéroport ; latInc et longInc
désignent la latitude et la longitude de l'incendie

distance <- function(latAero, longAero, latInc, longInc)
{
  # Conversion degrés->radian des coordonnées géographiques

  latAero = NISTdegTOradian(latAero)
  longAero = NISTdegTOradian(longAero)
  latInc = NISTdegTOradian(latInc)
  longInc = NISTdegTOradian(longInc)

```

```

Na <- 6378.0^2 / sqrt(6378.0^2*cos(latAero)^2 + 6356.0^2*sin(latAero))
Ni <- 6378.0^2 / sqrt(6378.0^2*cos(latInc)^2 + 6356.0^2*sin(latInc))
Xa = Na * cos(latAero) * cos(longAero)
Ya = Na * cos(latAero) * sin(longAero)
Za = (6356.0^2/6378.0^2) * Na * sin(latAero)

Xi = Ni * cos(latInc) * cos(longInc)
Yi = Ni * cos(latInc) * sin(longInc)
Zi = (6356.0^2/6378.0^2) * Ni * sin(latInc)

d = sqrt((Xi - Xa)^2 + (Yi - Ya)^2 + (Zi - Za)^2)
return (d)
}

# La fonction getValue retourne, pour un nom de variable (température, vitesse du vent,...) dont
la balise HTML est présente dans la page, la valeur numérique qui lui est associée.

# La variable html désigne le code HTML de la page et td le nom de la variable (« temperature »,
« precipitation »,...).

getValue <- function(html,td){
  indice = -1

  # On recherche la position de la balise contenant le nom de la variable dans le code HTML.
  indice = which(html==td)

  # La balise "Precipitation" apparaît deux fois dans les pages...
  if ((length(indice) >= 0) && (td=="Precipitation")){
    indice = indice[2]
  }

  chaine = ""

  if ((length(indice) > 0) && (indice >= 0)){

    # Lorsque la balise a été trouvée dans la page, on récupère la chaîne de caractères située
    dans la balise suivante qui correspond alors à la valeur numérique associée à la variable.

    # Rappelons que les données sont répertoriées dans un tableau, deux balises consécutives
    correspondent donc à deux cases consécutives d'une même ligne.

    chaine = html[indice+1]

    # on élimine les espaces contenus dans la chaîne.
    chaine = str_trim(chaine)
    ind = str_length(chaine)

    # On élimine les caractères associés à l'unité de la variable
    if (grepl("Temperature",td) !=0) {

```

```

ind = str_locate(chaine,"°C") }
if (grepl("Precipitation",td) !=0) {
  ind = str_locate(chaine,"mm") }
if (grepl("Wind",td) !=0) {
  ind = str_locate(chaine,"km/h") }
if (grepl("Humidity",td) == 0) {
  chaine = str_sub(chaine,1,ind[1]-1) }
chaine = str_trim(chaine)
}

# Enfin, on convertit la chaîne de caractère résultante en variable numérique
value = as.numeric(chaine)
return (value)
}

```

URL où se trouve la liste complète des aéroports dont on veut extraire les données météorologiques (cf. annexe A)

```
list.url<-http://www.wunderground.com/history/index.html?error=AMBIGUOUS&query=France&day=17
&month=8&year=2015&MR=1"
```

Liste des balises contenant les liens vers chacun des aéroports

```
node <- "#history-station-search-row li a"
```

Capture des liens vers les pages des aéroports

```
airports.list <- read_html(list.url) %>% html_nodes(node) %>% html_attr("href")
```

```
airports.names <- read_html(list.url) %>% html_nodes(node) %>% html_text
```

On retient parmi la liste des noms d'aéroports obtenus ceux qui sont présent dans la zone couverte par le réseau Prométhée et dont l'historique météo est disponible sur le site.

```
airports_list<-airports.names[c(4,5,23,29,50,58,60,64,90,93,96,109,111,117,120,129,130,134,135,
138, 147,159,164,168,171)]
```

On ajoute l'aéroport de Levaldigi (cf. 2.2)

```
airports_list[length(airports_list)+1] = "Levaldigi, Italy"
```

On calcule pour chaque aéroport ses coordonnées géographiques (latitude, longitude) à partir du nom de la ville sur laquelle ils sont situés

```
latlong <- geocode(airports_list)
```

```
aeroports = as.data.frame(airports_list)
```

```
aeroports$long = latlong$lon
```

```
aeroports$lat = latlong$lat
```

On associe à chaque aéroport de la liste la chaîne de caractère qui le caractérise dans l'URL associé à son historique (cf. 5.1)

```
airports.code<-c("LFKJ","LFCI","LFLW","LFKB","LFKC","LFKF","LFMK","LFBC","LFLS","LFTH","LFMI"
"LFMC","LFHP","LFLY","LFML","LFLQ","LFMT","LFMN","LFME","LFMO","LFMP","LFCR","LFMH","LF
MY","LFKS","LIMZ","LFBO")
```

```
aeroports$code = airports.code
```

On récupère la liste des incendies de 2002 à 2014.

```
feux_2002_2014=read.csv("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/feux_2002_2014
.csv", header=FALSE, sep=";")
```

On conserve les colonnes suivantes : année, code postal, heure et date de l'alerte incendie, surface brûlée au début et à la fin de l'intervention, carreau DFCI, type de forêt, distance des habitations, heure et date de début d'intervention, heure et date de fin d'intervention

```
feux = feux_2002_2014[,-c(2,3,4,6,7,8,10)]
```

```
names(feux)[1]="ANNEE"
```

```
names(feux)[2]="CP"
```

```
names(feux)[3]="DATE"
```

```
names(feux)[4]="SURFACE"
```

```
names(feux)[5]="DFCI"
```

```
names(feux)[6]="TYPE_FORET"
```

```
names(feux)[7]="DIST_HAB"
```

```
names(feux)[8]="DEB_INT"
```

```
names(feux)[9]="FIN_INT"
```

```
names(feux)[10]="SURFACE_DEB"
```

On convertit les dates au format date de R

```
feux$DATE_HEURE = parse_date_time(feux$DATE,'%d/%m/%y %h:%M')
```

```
feux$DEB_INT = parse_date_time(feux$DEB_INT,'%d/%m/%y %h:%M')
```

```
feux$FIN_INT = parse_date_time(feux$FIN_INT,'%d/%m/%y %h:%M')
```

On calcule le délai entre l'alerte et le début de l'intervention (en minutes)

```
feux$DELAI = round(difftime(feux$DEB_INT , feux$DATE_HEURE, units="mins"))
```

On calcule la durée de l'intervention (en minutes)

```
feux$DUREE = round(difftime(feux$FIN_INT , feux$DATE_HEURE, units="mins"))
```

On rajoute un '0' devant les codes INSEE à 4 chiffres ; ceci concerne les départements dont le numéro INSEE est inférieur à 10 (Alpes Maritimes (06), Alpes de Haute Provence (04),...).

Cette modification est indispensable pour géocoder l'incendie.

```
feux$CP = str_c("0",feux$CP)
```

```
feux$CP = str_sub(feux$CP,-5)
```

Création de la variable département

```
feux$dep = NULL  
code_dep = c("83","2A","26","30","84","04","06","11","2B","13","07","48","05","66","34")  
no_dep = c(1:15)  
for (i in 1:length(feux$CP))  
{  
  feux$dep[i] = no_dep[which(code_dep == str_sub(feux$CP[i],1,2))]  
}
```

Ouverture du fichier contenant l'ensemble des informations démographiques et géographiques des communes de France : population, superficie et altitude (obtenu sur le site officiel de l'INSEE)

```
donnees_Communes=read.csv("C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/communes  
Simples.csv", header=TRUE, sep=";")  
donnees_Communes$Population = donnees_Communes$Population * 1000.0  
donnees_Communes$Superficie = donnees_Communes$Superficie * 0.01  
donnees_Communes$densite = donnees_Communes$Population / donnees_Communes$Superficie  
donnees_Communes$Code = str_c("0",donnees_Communes$Code)  
donnees_Communes$Code = str_sub(donnees_Communes$Code,-5)
```

On rattache à chaque feu les données de la commune sur laquelle il est apparu

```
Densite = NULL  
Population = NULL  
Superficie = NULL  
for (i in 1:length(feux$DATE_HEURE))  
{  
  if (feux$CP[i] == "13055")  
  {  
    Densite[i] = 400  
    Population[i] = 100000  
    Superficie [i] = 0  
  }  
  else if (feux$CP[i] == "83999")  
  {  
    Densite[i] = 50.0  
    Population[i] = 1000  
    Superficie [i] = 0  
  }  
}
```

```

}
else
{
  Densite[i] = floor(donnees_Communes$densite[which(donnees_Communes$Code == feux$CP[i])])
  Population[i] = donnees_Communes$Population[which(donnees_Communes$Code == feux$CP[i])]
  Superficie[i] = donnees_Communes$Superficie[which(donnees_Communes$Code == feux$CP[i])]
}
}
feux$densitePopulation = Densite
feux$Population = Population
feux$Superficie_Commune = Superficie

```

Pour chaque incendie, nous recherchons maintenant l'aéroport le plus proche. On utilise pour cela la fonction *distance* définie plus haut ; on calcule la distance entre l'incendie et chaque aéroport de la liste et on retient l'aéroport dont la distance à l'incendie est la plus faible.

```

aeroportFeux <- NULL
code <- NULL
for (i in 1:length(feux$LAT))
{
  d = 1000000;
  for (j in 1:length(aeroports$lat))
  {
    dist = distance(aeroports$lat[j], aeroport$long[j], feux$latitude[i], feux$longitude[i])
    if (dist < d)
    {
      d = dist
      aeroportFeux[i] = as.character(aeroports$airports_list[j])
      code[i] = as.character(aeroports$code[j])
    }
  }
}

```

On ajoute à la data frame FEUX le code de l'aéroport identifiant l'URL de son historique ainsi que le nom de l'aéroport.

```

feux$code = code
feux$aeroport = aeroportFeux

```

Initialisation des tableaux contenant les données météo de chaque incendie (Température moyenne du jour de l'alerte incendie, température max, taux de précipitation, taux d'humidité moyen, taux d'humidité max, vitesse du vent moyenne et vitesse du vent max)

```

Tmoy <- NULL
Tmax <- NULL
Prec <- NULL
Hmoy <- NULL
Hmax <- NULL
Vmoy <- NULL
Vmax <- NULL

url_base = "https://www.wunderground.com/history/airport"
url_end = "DailyHistory.html"

# Boucle sur l'ensemble des incendies dans laquelle nous irons récupérer pour chacun d'entre eux les données météo disponibles sur le site wunderground.com

for (i in 1:length(feux$ANNEE)) {
  url_complete = ""

  # La date des incendies est écrite sous la forme JJ/MM/AAAA HH:MM:SS ; on ôte donc de la chaîne de caractère la partie droite associée à l'heure de l'alerte incendie. On enlève ensuite les '/' de la partie gauche dans la mesure où ils ne figurent pas dans l'URL des historiques du site wunderground

  date = str_split(feux $DATE[i], " ");
  date2 = str_split(date[[1]][1], "/");

  # On peut alors écrire l'URL complète fournissant les données météo de l'aéroport associé à l'incendie au jour de l'alerte en concaténant les éléments nécessaires dans l'ordre suivant : code de l'aéroport, puis jour, mois et année de l'incendie

  url_complete=paste(url_base,feux$code[i],date2[[1]][3],date2[[1]][2],date2[[1]][1],url_end,sep="/");

  # On lit le code HTML de la page associée

  page <- read_html(url_complete)

  # On récupère toutes les balises associées à une case de tableau (balise de type « td » en langage HTML)

  temp <- html_nodes(page, xpath="//table//td") %>% html_text()

  # On récupère les données météo parmi ces balises à l'aide de la fonction getValue définie plus haut

  Tmoy[i] = getValue(temp,"Mean Temperature")
  Tmax[i] = getValue(temp,"Max Temperature")
  Prec[i] = getValue(temp,"Precipitation")
  Hmoy[i] = getValue(temp,"Average Humidity")
  Hmax[i] = getValue(temp,"Maximum Humidity")
  Vmoy[i] = getValue(temp,"Wind Speed")
  Vmax[i] = getValue(temp,"Max Wind Speed")
}

# On ajoute les colonnes ainsi obtenues à la data frame feux

```

```

feux$Tmoy = Tmoy
feux$Tmax = Tmax
feux$Prec = Prec
feux$Hmoy = Hmoy
feux$Hmax = Hmax
feux$Vmoy = Vmoy
feux$Vmax = Vmax

# Tri des données

# On ôte les lignes dont la valeur du taux de précipitation est indéterminée.
vec_na = which(is.na(feux$Prec))

feux <- feux[-vec_na,]

# On ôte les lignes dont la valeur de la température moyenne est indéterminée.
vec_na = which(is.na(feux$Tmoy))

feux <- feux[-vec_na,]

# On ôte les feux de forêt dont la surface brûlée est inférieure à un hectare.
vec_feu = which(feux$SURFACE <= 10000)

feux <- feux[-vec_feu,]

# On ôte les feux apparaissant en dehors de la période estivale.
month = month(feux$DATE)

feux$month = month

vec_ete = which(feux$month %in% c(1:6,10:12))

feux <- feux[-vec_ete,]

# Sauvegarde de la data frame feux avec les coordonnées géographiques des incendies et les données météo du jour des alertes incendies

write.csv(feux,file="C:/feux_aeroports_avecMeteo.csv",row.names = FALSE, na="")

```

D3. Elaboration du modèle SVM basé sur la météo uniquement

```
library(dplyr) # Librairie de manipulation de dataframes
```

```
library(kernlab) # Librairie contenant l'algorithme SVM
```

```
library(pheatmap) # Affichage des heatmaps
```

```
library(ROCR) # Librairie de calcul et visualisation de résultats de prédictions
```


On lit les données préparées précédemment

```
feux=read.csv("C:/feux_aeroports_avecMeteo.csv ", header=TRUE, sep=";")
```

On répartit les incendies en 2 groupes Y=1 ou Y=-1 selon que la surface brûlée soit respectivement supérieure ou inférieure à 5 hectares (les données d'entrée sont en m²)

```
Y = feux$SURFACE
```

```
Y[feux$SURFACE > 50000] = 1
```

```
Y[feux$SURFACE <= 50000] = -1
```

```
feux$Y = Y
```

On découpe les données en 2 : les données d'apprentissage du modèle correspondent aux incendies survenus durant les années 2002 à 2009 incluses ; les incendies survenus durant les années 2010 à 2014 incluses constituent les données de validation du modèle

```
feux_train <- filter(feux, ANNEE %in% c(2002 :2009))
```

```
feux_test <- filter(feux, ANNEE %in% c(2010 :2014))
```

On utilise pour variables de notre modèle la température moyenne, le taux moyen d'humidité et la vitesse moyenne du vent du jour de l'alerte incendie.

```
meteo_train <- feux_train[,c("Tmoy", "Hmoy", "Vmoy")]
```

```
meteo_test <- feux_test[,c("Tmoy", "Hmoy", "Vmoy")]
```

```
Y_train <- feux_train[,c("Y")]
```

```
Y_test <- feux_test[,c("Y")]
```

Tracé de la répartition des 2 groupes d'incendies (surface brûlée supérieure ou inférieure à 5 hectares). Les ronds verts correspondent aux feux ayant brûlé entre 1 et 5 hectares et les ronds rouges les feux ayant brûlé plus de 5 hectares.

```
YY = Y_train ; YY[Y_train==1] = 2
```

```
plot(meteo_train, main = "INCENDIES",
```

```
      pch = c(21, 22)[YY],
```

```
      bg = c("red", "green3")[YY]
```

```
)
```

On fixe une plage de variation pour les paramètres γ et σ afin de déterminer leur valeur optimale pour notre modèle (le vecteur C correspond à la variable γ) :

```
C = seq(0.01, 50, length = 100)
```

```

sigma = seq(0.01, 50, length = 100)

RESULTS = matrix(0, length(C), length(sigma))

for (i in 1 : length(C)){
  for (j in 1 : length(sigma)){

# Calcul du modèle selon les SVM à noyau gaussien pour chaque valeur de  $\gamma$  et  $\sigma$ 

    res.svm = ksvm(as.matrix(meteo_train), Y_train,

      type = "C-svc",

      kernel = "rbfdot", # Type de noyau

      kpar = list(sigma = sigma[j]), # paramètres du noyau

      cross = 7, # nombre de tests utilisés pour la validation croisée

      C=C[i]) # Valeur de  $\gamma$ 

# On stocke chaque taux d'erreur associé au couple ( $\gamma, \sigma$ ) dans un tableau.

    RESULTS[i, j] = res.svm@cross

  }
}

# Visualisation des taux d'erreurs en fonction de  $\gamma$  et  $\sigma$  dans une heatmap

pheatmap(RERESULTS, cluster_rows = FALSE, cluster_cols = FALSE)

# On récupère et affiche les indices du tableau donnant le couple ( $\gamma, \sigma$ ) optimal

re = which(RERESULTS == min(RERESULTS), arr.ind=TRUE)

re

# On retient le modèle associé à ce couple ( $\sigma = 21$  ,  $\gamma = 1$ )

resfinal = ksvm(as.matrix(meteo_train), Y_train, kernel = "rbfdot",

  type = "C-svc",

  kpar = list(sigma = sigma[21]),

  cross=7, C = C[1])

# On applique le modèle obtenu aux données de validation

Yhat = predict(resfinal, as.matrix(meteo_test), type = "response")

```

On en déduit la matrice de confusion en comparant avec les données de validation réelles

```
CONF = table(Y_test, Yhat)
```

```
CONF
```

D4. Modèle prenant en compte les variables environnementales

```
library(dplyr)
```

```
library(ROCR)
```

```
library(lubridate)
```

```
library(stringr)
```

```
library(randomForest)
```

```
library(kernlab)
```

Lecture du fichier contenant l'ensemble des données liées aux incendies (cf. annexe D2 sur la préparation des données) : météo, données liées aux incendies (date, délai et durée d'intervention, surface brûlée, type de forêt, démographie,...)

```
feuxRandom=read.csv(file= "C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/feux_complet_avecMeteo.csv", na="")
```

Sauvegarde du data frame qui vient d'être lu

```
feuxRandom_sav = feuxRandom
```

```
head(feuxRandom_sav)
```

On ne garde que les lignes où le type végétation est renseigné

```
vec_na = which(is.na(feuxRandom_sav$TYPE_FORET))
```

```
feuxRandom_sav <- feuxRandom_sav[-vec_na,]
```

On crée 6 catégories de population

```
feuxRandom_sav$rural=feuxRandom_sav$Population
```

```
feuxRandom_sav$rural[feuxRandom_sav$Population <= 100] = 1
```

```
feuxRandom_sav$rural[feuxRandom_sav$Population>100 & feuxRandom_sav$Population <= 500] = 2
```

```
feuxRandom_sav$rural[feuxRandom_sav$Population>500 & feuxRandom_sav$Population<=1500] = 3
```

```
feuxRandom_sav$rural[feuxRandom_sav$Population>1500& feuxRandom_sav$Population<=3000]=4
```

```
feuxRandom_sav$rural[feuxRandom_sav$Population>3000& feuxRandom_sav$Population<=6000]=5
```

```
feuxRandom_sav$rural[feuxRandom_sav$Population > 6000 ] = 6
```

On réalise l'analyse de variance par rapport à la démographie et à la végétation

```
res.aov = aov(feuxRandom_sav$SURFACE~feuxRandom_sav$rural)
```

```
summary(res.aov)
```

```
res.aov = aov(feuxRandom_sav$SURFACE~feuxRandom_sav$TYPE_FORET)
```

```
summary(res.aov)
```

Affichage des box-plot

```
boxplot(feuxRandom_sav$SURFACE/10000.0~feuxRandom_sav$TYPE_FORET,col="red",ylim=c(0,30), names = c("Maquis","Taillis","Fut. Feuillus", "Fut. résineux", "Fut. mixtes","Reboisement"))
```

```
boxplot(feuxRandom_sav$SURFACE/10000.0~feuxRandom_sav$rural,col="red",ylim=c(0,20), names=c("Catégorie 1", "Catégorie 2", "Catégorie 3", "Catégorie 4", "Catégorie 5", "Catégorie 6"))
```

Variable à prédire : Y = 1 si surface brûlée > 5 hectares et Y = -1 sinon

```
Y = feuxRandom_sav$SURFACE
```

```
Y[feuxRandom_sav$SURFACE > 50000] = 1
```

```
Y[feuxRandom_sav$SURFACE <= 50000] = -1
```

```
feuxRandom_sav$Y = Y
```

On initialise le générateur de nombres aléatoires (indispensable pour les random forest)

```
set.seed(123)
```

Tracé de l'histogramme indiquant la répartition du nombre d'incendies en fonction du délai d'intervention

```
hist(as.integer(feuxRandom_sav$DELA), breaks = c(0:120), xlab = "Délai d'intervention (en minutes)", ylab = "Fréquence")
```

On crée des variables de type FACTEUR pour les variables qualitatives telles que le département, le type de forêt, le nombre d'habitants... nécessaire pour les random forest

```
feuxRandom_sav$Yb = as.factor(feuxRandom_sav$Y)
```

```
feuxRandom_sav$dep1 = as.factor(feuxRandom_sav$dep)
```

```
feuxRandom_sav$TYPE_FORET1 = as.factor(feuxRandom_sav$TYPE_FORET)
```

```
feuxRandom_sav$rural1 = as.factor(feuxRandom_sav$rural)
```

Sélection des variables intervenant dans le modèle : T°C, Vent, Humidité, précipitations, département, démographie, superficie de la commune, végétation

```
feuxRandom_sav2 <- select(feuxRandom_sav, Tmoy, Vmoy, Prec, Yb, rural1, Superficie_Commune, dep1, TYPE_FORET1)
```

Application de l'algorithme à la dataframe résultante

```
fit <- randomForest(formula = Yb ~ ., # On prédit la variable Yb à partir de l'ensemble des prédictors retenus ci-dessus
```

```
data = feuxRandom_sav2, # data frame contenant les prédictors
```

```
na.action = na.roughfix, # On remplace les valeurs non déterminées d'un prédictor par sa valeur moyenne
```

```
nntree= 5000, # Nombre d'arbres aléatoires
```

```
mtry = 3) # Nombre de prédictors retenus pour chaque arbre (par défaut sqrt(n), n étant le nombre total de prédictors)
```

Visualisation de la matrice de confusion

```
print(fit)
```

Visualisation de l'importance des différentes variables

```
varImpPlot(fit, labels=c("T moyenne", "Vitesse vent moyenne", "Précipitations", "Population", "Superficie commune", "Département", "Type de forêt"))
```

Tracé du taux d'erreur en fonction du nombre d'arbres

```
plot(fit$err.rate[, 1], type = "l", xlab = "nombre d'arbres", ylab = "erreur OOB")
```

Passage aux SVM avec les mêmes données d'entrée

On coupe les données en données d'entraînement et en données de validation

```
feux_train<-filter(feuxRandom_sav,ANNEE%in%c(2002,2003,2004,2005,2006,2007,2008,2009,2010))
```

```
feux_test <- filter(feuxRandom_sav, ANNEE %in% c(2001,2013,2014,2011,2012))
```

On transforme pour chacun de ces 2 échantillons les variables qualitatives en variables binaires : indispensable pour les SVM avec variables qualitatives

```
rural_train = acm.disjonctif(data.frame(feux_train$rural))
```

```
foret_train = acm.disjonctif(data.frame(feux_train$TYPE_FORET))
```

```
rural_test = acm.disjonctif(data.frame(feux_test$rural))
```

```
foret_test = acm.disjonctif(data.frame(feux_test$TYPE_FORET))
```

```
dep_train = acm.disjonctif(data.frame(feux_train$dep))
```

```
dep_test = acm.disjonctif(data.frame(feux_test$dep))
```

On regroupe pour les 2 échantillons (entraînement et validation) ces variables binaires avec les données quantitatives intervenant dans le modèle, notamment la météo

```
feux_train2=data.frame(dep_train,rural_train,foret_train,dep_train,feux_train$Tmoy,feux_train$Hmoy,feux_train$Vmoy,feux_train$SUPERFICIE_COMMUNE)
```

```
feux_test2=data.frame(dep_test,rural_test,foret_test,dep_test,feux_test$Tmoy,feux_test$Hmoy,feux_test$Vmoy,feux_test$SUPERFICIE_COMMUNE)
```

```
Y_train <- select(feux_train, Y)
```

```
Y_test <- select(feux_test, Y)
```

On fait varier C et sigma entre 0 et 50, comme précédemment

```
C = seq(0.01, 50, length = 100)
```

```
sigma = seq(0.01, 50, length = 100)
```

```
RESULTS = matrix(0, length(C), length(sigma))
```

```
for (i in 1 : length(C)){
```

```
  for (j in 1 : length(sigma)){
```

```
    res.svm = ksvm(as.matrix(feux_train2), Y_train,
```

```
      type = "C-svc",
```

```
      kernel = "rbfdot",
```

```

        kpar = list(sigma = sigma[j]),
        cross = 7,
        C=C[i])

    RESULTS[i, j] = res.svm@cross
}
}

# Visualisation des taux d'erreurs
pheatmap(RESULTS, cluster_rows = FALSE, cluster_cols = FALSE)

# On récupère et affiche les indices du tableau donnant le couple ( $\gamma, \sigma$ ) optimal
re = which(RESULTS == min(RESULTS), arr.ind=TRUE)

re

# On retient le modèle associé au couple ( $\gamma, \sigma$ ) optimal
resfinal = ksvm(as.matrix(feux_train2), Y_train, kernel = "rbfdot",
               type = "C-svc",
               kpar = list(sigma = sigma[37]),
               C = C[71],
               cross=7)

# Matrice de confusion obtenues à partir des prévisions du modèle appliqué aux données de validation
Yhat = predict(resfinal, feux_test2, type = "response")
CONF = table(as.matrix(Y_test), Yhat)
CONF

```

D5. Corrélations entre variables liées à l'intervention des secours

```

library(dplyr)

# Lecture du fichier contenant l'ensemble des données liées aux incendies
feuxRandom = read.csv(file="C:/Users/MVE05912/Desktop/PROJET_FEUX_FORETS/feux_2002_2014_distances_demographie_last_complet_last.csv", na="")

# Tracé de la relation :  $\ln(\text{surface finale}) \sim \ln(\text{surface initiale})$ 
plot(y = log(feuxRandom_sav$SURFACE/10000.0), x = log(feuxRandom_sav$SURFACE_DEB/10000.0), pch=21, bg="red", xlab="Surface brûlée à l'arrivée des secours", ylab = "Surface brûlée après intervention")

# Régression linéaire entre  $\ln(\text{surface finale})$  et  $\ln(\text{surface initiale})$ 
res.lm = lm(log(feuxRandom_sav$SURFACE/10000.0) ~ log(feuxRandom_sav$SURFACE_DEB/10000.0))
summary(res.lm)

```

Prédiction du modèle sur les données

```
pred = predict(res.lm, interval="prediction")
```

On affiche la droite de régression

```
abline(res.lm)
```

On affiche l'intervalle de confiance de la prediction

```
lines(x = log(feuxRandom_sav$SURFACE_DEB/10000.0),y=pred[,2], col="green")
```

```
lines(x = log(feuxRandom_sav$SURFACE_DEB/10000.0),y=pred[,3], col="green")
```

Calcul des coefficients de corrélation entre la surface au début de l'intervention, la surface finale, le délai d'arrivée des secours et la durée de l'intervention

```
y = data.frame(log(feuxRandom_sav$SURFACE), log(feuxRandom_sav$SURFACE_DEB),  
log(feuxRandom_sav$DELA), log(feuxRandom_sav$DUREE))
```

```
colnames(y) = c("Surface Finale", "Surface initiale", "Délai", "Durée")
```

```
round(cor(y),3)
```

E. NOMENCLATURE DE L'ORIGINE DES FEUX DE FORET UTILISEE PAR LE RESEAU PROMETHEE

