INSA – MS ESD Année 2019-2020
BRUNEVAL Guillaume
ROUFFIAC Jean-Eudes

STAT 1 : Sélection de variables et validation de modèles

Rapport de TP n°1

Titre : *Etude descriptive des types de couverture forestière*

1 Contextualisation

Le jeu de données de ce TP provient d'une étude américaine sur la couverture forestière dans le parc national Roosevelt, dans le nord du Colorado.

Ce jeu de données a été constitué en août 1998 à partir de relevés effectués les années précédentes par Jock A.Blackard, du "College of Natural Resources" de l'université du Colorado, ainsi que le Dr Denis J. Dean, professeur à "School of Economic, Political and Policy Sciences" de Richardson.

Ces données consistent en 581012 parcelles de terrain observées dans quatre zones différentes du parc national Roosevelt. Ces quatre zones ont été choisies parce qu'elles présentaient relativement peu de trace de l'intervention humaine. L'objectif étant de se rapprocher au maximum de l'étude d'une forêt "primaire".

Pour chaque observation, ces chercheurs ont relevé 13 données cartographiques : l'altitude, l'orientation, la pente, les distances horizontales et verticales à un point d'eau, la distance à une route, les mesures d'ombrage à 9h, midi et 15h, la distance à un feu, la zone dans laquelle la parcelle observée se situe, le type de sol et enfin l'espèce d'arbre présente sur la parcelle.

Les données ont ensuite donné lieu à trois études entre 1998 et 2000. L'objectif des chercheurs était initialement d'utiliser ces données pour entraîner un modèle capable de pouvoir prédire le type de couverture forestière présente sur une parcelle donnée du parc.

A l'origine de ces études, on trouve les agences fédérales américaines qui étaient dans le besoin d'un inventaire précis des ressources naturelles pour orienter leurs politiques de gestion des terres. Le type de couverture forestière est l'une des caractéristiques les plus fondamentales enregistrées dans ces inventaires.

Les auteurs ont alors appliqué des modèles statistiques de classification, des réseaux de neurones en l'occurrence, afin de prédire le type de couverture des forêts dans les quatre zones, ce qui a fait gagner beaucoup de temps aux agences fédérales.

Notre objectif pour ce TP est d'effectuer une première analyse exploratoire de ces données, c'est à dire d'effectuer une étude descriptive sur l'ensemble des données en vue d'une éventuelle étape de modélisation ultérieure.

2 Description des données

Nous allons dans cette partie présenter et décrire statistiquement les données à disposition en les identifiant et en mettant en évidence leurs caractéristiques.

Chaque observation correspond à une parcelle forestière de 30 mètres sur 30 mètres située dans une des quatre "zones sauvages" américaines prises en compte dans cette étude. Toutes les données sont complètes, c'est à dire qu'il n'y a aucune valeur manquante dans le jeu de données.

Nos 13 variables se décomposent en :

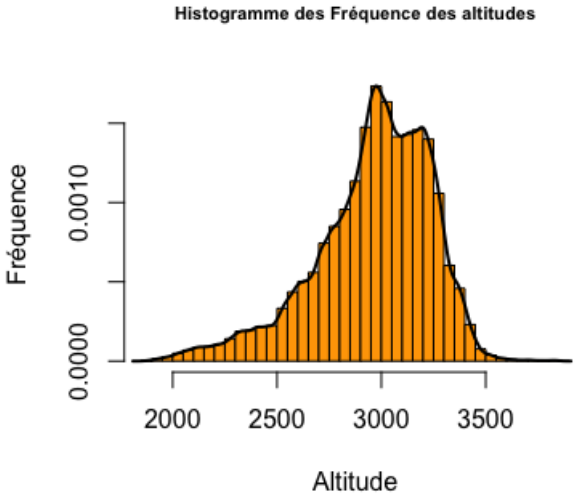
- 10 variables quantitatives : Altitude, Orientation, Pente, Distance_horizontale_point_eau, Distance_verticale_point_eau, Distance_horizontale_route, Ombrage_9am,Ombrage_midi, Ombrage_3pm, Distance_horizontale_feu
- 3 variables qualitatives : Zone sauvage, type de sol, et type de couverture.

Altitude Correspond à l'altitude de la parcelle, en mètres, de la zone.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Altitude	1859	2809	2996	2959	3163	3858	278	-0,82	3,75

FIGURE 1 – Indicateurs statistiques de la variable "Altitude"

Ces indicateurs nous donnent une information sur la manière dont sont réparties les altitudes. On remarque que les altitudes sont hautes, 75% des valeurs étant supérieures à 2809m. La moyenne est légèrement inférieure à la médiane, ce qui indique une légère distribution asymétrique à gauche. Ceci est confirmé par le skewness négatif et l'histogramme qui montre une distribution étalée vers la gauche.



Orientation Correspond à l'orientation de la parcelle en degré azimut (le nord est pris comme de degré 0).

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Orientation	0.0	58.0	127.0	155.7	260.0	360.0	112	0.40	1.78

FIGURE 2 – Indicateurs statistiques de la variable "Orientation"

Les orientations sont très reparties. La valeur de l'écart type de 112, pour une médiane de 155 confirme cela.

Pente Correspond à la pente en degré de la parcelle.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Pente	0.0	9.0	13.0	14.1	18.0	66.0	7.48	0.78	3.58

FIGURE 3 – Indicateurs statistiques de la variable "Pente"

50% des valeurs de pente sont comprises entre 9 et 18 degrés. La médiane est à 14 degrés alors que la moyenne est à 13 degrés. Tout ceci indique que les valeurs sont sensiblement regroupées autour de la moyenne, avec une légère asymétrie à droite, ce que confirme le skewness de 0.78.

Distance_horizontale_point_eau Correspond à la distance horizontale en mètres entre la parcelle et le point d'eau le plus proche.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Distance horizontale point d'eau	0.0	108.0	218.0	269.4	384.0	1397.0	212.54	1.14	4.36

FIGURE 4 – Indicateurs statistiques de la variable "Distance_horizontale_point_eau"

La majorité des parcelles sont proches d'un point d'eau puisque 75% des valeurs sont inférieures à 384m.

Distance_verticale_point_eau Correspond à la distance verticale en mètres entre la parcelle et le point d'eau le plus proche.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Distance verticale point d'eau	-173.00	7.00	30.00	46.42	69.00	601.00	58.29	1.79	8.25

FIGURE 5 – Indicateurs statistiques de la variable "Distance_verticale_point_eau"

50% des parcelles sont entre 7 et 69m d'un point d'eau. Mais il y a de grandes dispersions puisque le maximum est de 601m et le minimum de -173m.

Distance_horizontale_route Correspond à la distance horizontale minimale en mètres entre la parcelle et une route.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Distance horizontale à une route	0	1106	1997	2350	3328	7117	1559	0.71	2.61

FIGURE 6 – Indicateurs statistiques de la variable "Distance_horizontale_route"

La moyenne est de de 2350m et la médiane à 1997m. Cela indique une distribution particulièrement asymétrique à droite.

Ombrage_9am Correspond à la valeur d'ombrage à 9 heure du matin sur la parcelle pour un jour de référence qui est le solstice d'été. C'est une notion peu connue des non spécialistes que l'on peut expliciter ici. Il s'agit en fait d'un indice compris entre 1 et 255 calculé qui mesure en quelque sorte la quantité d'ombre sur la parcelle. C'est à dire que moins la parcelle recevra de luminosité, plus son indice sera haut.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Ombrage à 9h	0.0	198.0	218.0	212.1	231.0	254.0	26.76	-1.18	4.87

FIGURE 7 – Indicateurs statistiques de la variable "Ombrage_9am"

75% des valeurs sont supérieures à 198, les parcelles ne reçoivent globalement pas beaucoup de luminosité à 9h du matin.

Ombrage_midi Correspond à la valeur d'ombrage à midi sur la parcelle pour un jour de référence qui est le solstice d'été.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Ombrage à 12h	0.0	213.0	226.0	223.3	237.0	254.0	19.76	-1.06	5.06

FIGURE 8 – Indicateurs statistiques de la variable "Ombrage_midi"

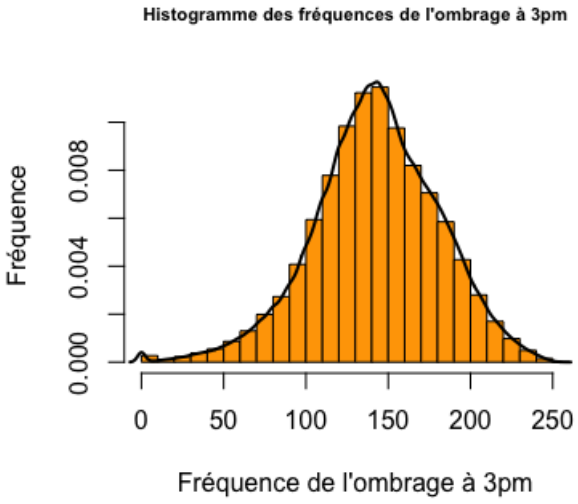
On remarque qu'à 12h, les parcelles sont encore plus à l'ombre qu'à 9h. En effet, la moyenne est à 223 et la médiane à 226, avec un écart-type de 19 ce qui indique une faible dispersion autour de la moyenne.

Ombrage_3pm Correspond à la valeur d'ombrage à 3 heure de l'après-midi sur la parcelle pour un jour de référence qui est le solstice d'été.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Ombrage à 15h	0.0	119.0	143.0	142.5	168.0	254.0	38.27	-0.27	3.39

FIGURE 9 – Indicateurs statistiques de la variable "Ombrage_3pm"

A 15h, il y a moins d'ombre sur les parcelles qu'à 9h et à midi, ce qui est logique. Les valeurs de la variable "Ombrage_3pm" semblent se rapprocher d'une distribution gaussienne. En effet, la moyenne et la médiane sont égales, le kurtosis est proche de 3. De plus, nous pouvons voir la distribution sur l'histogramme qui se rapproche d'une distribution gaussienne. On remarque que la distribution s'étale légèrement à gauche, ce que nous indique aussi le skewness négatif.



Distance_horizontale_feu Correspond à la distance horizontale en mètres de la parcelle à un point ayant déjà été un départ de feu.

Les distances à un point ayant déjà été un départ de feu sont pour 75% d'entre elles inférieures à 2550m. Des valeurs entre 2550m et 7173m s'étalent ensuite.

	Min.	1st Qu	Median	Mean	3rd Qu.	Max	écart-type	skewness	kurtosis
Distance horizontale à un feu	0	1024	1710	1980	2550	7173	1324	1.28	4.64

FIGURE 10 – Indicateurs statistiques de la variable "Distance_horizontale_feu"

Zone_sauvage Correspond à la zone dans laquelle se trouve la parcelle. Ce sont des zones situées aux Etats-Unis, appelées "Rawah", "Neota", "Comanche Peak" et "Cache la Poudre".

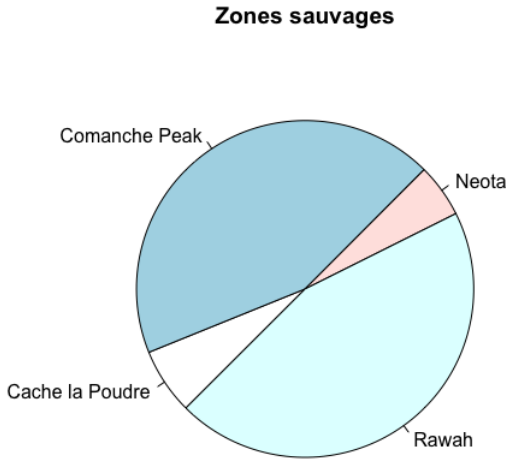


FIGURE 11 – Répartition de la variable Zone_sauvage

On remarque que la grande majorité des parcelles sont localisées dans les zones "Comanche Peak" et "Rawah".

Sol Correspond au type de sol présent sur la parcelle. C'est une variable qualitative codée numériquement puisqu'il y a quarante types de sol. Les types de sols sont codés de 1 à 40 à partir d'un index que nous avons décidé de ne pas reproduire afin d'optimiser le taux de contenu pertinent dans notre étude mais qui est consultable sur le fichier "info" joint au dataset.

1	2	3	4	5	6	7	8	9	10
3031	7525	4823	12396	1597	6575	105	179	1147	32634
11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
12410	29971	17431	599	3	2845	3422	1899	4021	9259
21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
838	33373	57752	21278	474	2589	1086	946	115247	30170
31	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
25666	52519	45154	1611	1891	119	298	15573	13806	8750

FIGURE 12 – Répartition de la variable "Sol"

Les répartitions des différents types de sol ne sont pas du tout équitables. En effet, certains types de sol sont présents qu'une centaine de fois contre plus de 115 000 pour le type de sol 29 par exemple.

Espèce Correspond à l'essence d'arbre majoritairement présente sur la parcelle, on parle de couverture forestière de la parcelle. On a décelé sur les parcelles sept espèces différentes qui sont les sept modalités de notre variable "Espèce" : Spruce/Fir, Lodgepole Pine, Ponderosa Pine, Cottonwood/Willow, Aspen, Douglas-fir et Krummholz.

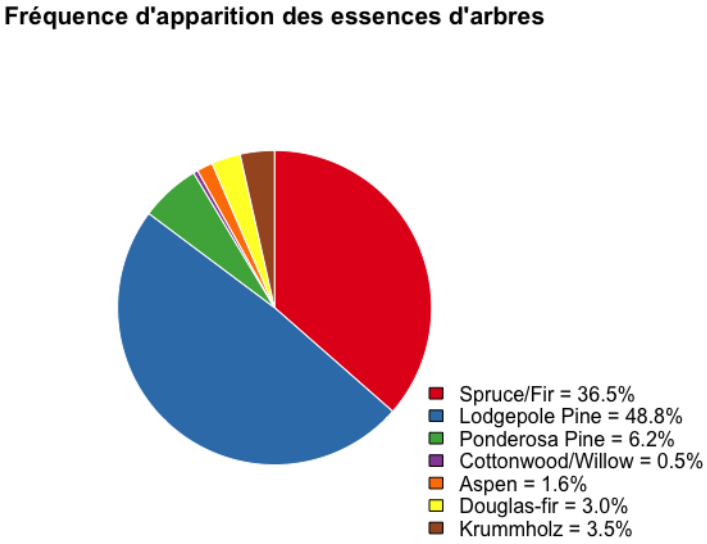


FIGURE 13 – Répartition de la variable "Espèce"

Le diagramme en secteur ci-dessus indique la répartition des différentes essences d'arbres. On peut voir que les espèces "Spruce/Fir" et "Lodgepole Pine" sont largement représentées puisqu'à elles seules, elles représentent 85% des données.

3 Liens entre les données

Afin de décrire au mieux le jeu de données, nous pouvons mettre en relation les variables pour regarder si il existe des liens entre elles.

Méthodologie : Nous avons tenté de comparer de manière systématique les variable deux à deux. Pour ce faire, nous avons :

- tracé les nuages de points entre chaque paire de variables quantitatives.
- tracé les boîtes à moustaches comparées entre chaque variable qualitative et chaque variable quantitative.
- construit les tables de contingences et calculé le coefficient de Kendall entre chaque paire de variables qualitatives.

De cette visualisation, il ressort plusieurs enseignements :

- il est à priori peu pertinent d'étudier les liens entre variables quantitatives. Premièrement, l'importance du nombre d'observations rend les nuages de points peu lisible. Mais nous aurions pu procéder à une ou plusieurs réduction aléatoire du jeu de données pour nos tracés afin de pouvoir se rendre compte simplement de la pertinence d'une étude plus approfondie. Cependant, la forme des nuages de points ne laisse pas augurer d'une quelconque corrélation linéaire ou même de lien quadratique entre les variables. La seul nuage de point qui pourrait susciter l'intérêt est celui de la distance horizontale en fonction de la distance verticale à un point d'eau. Lien à priori évident sur lequel on peut s'abstenir d'autre commentaire.
- l'étude des relations entre la zone étudiée et les autres variables semble en revanche plus fructueuse. Elle permettrait d'établir une caractérisation des zones en fonction de variables pertinentes.
- l'étude des relations entre l'espèce d'arbre présente sur la parcelle et les autres variables semble également digne d'intérêt. Ceci aurait tendance à orienter une future étape de modélisation vers une classification en fonction des sept essences d'arbres.

Caractérisation des zones

Dans un premier temps, nous souhaitons connaître les caractéristiques des quatre zones forestières.

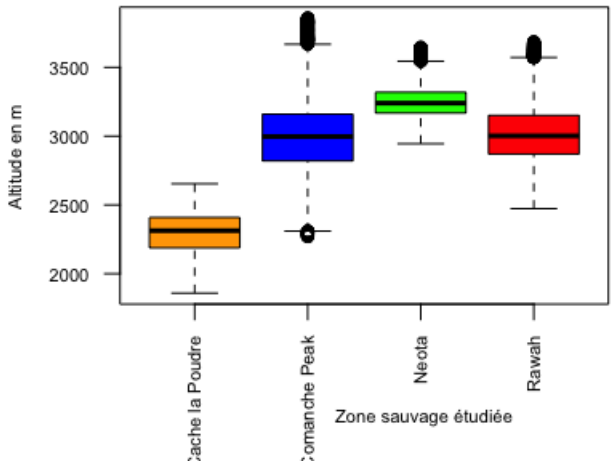


FIGURE 14 – Altitude en fonction de la zone forestière

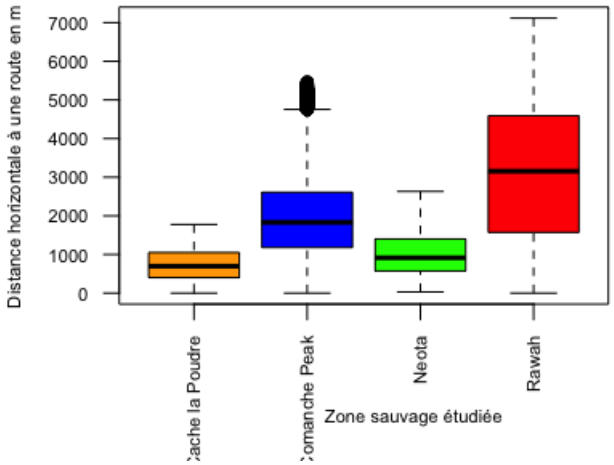


FIGURE 15 – Distance horizontale à une route en fonction de la zone forestière

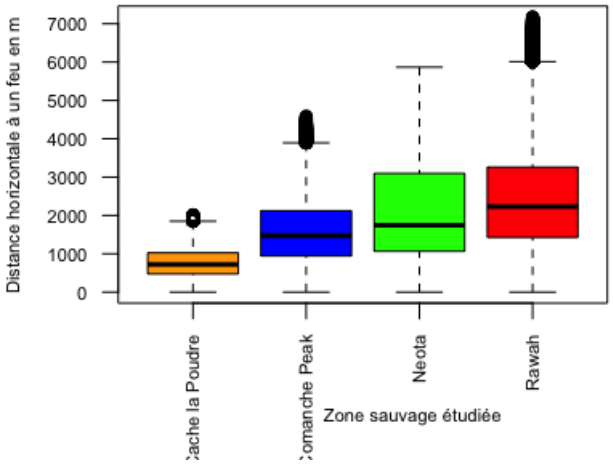


FIGURE 16 – Distance à un point départ de feu en fonction de la zone forestière

	Spruce/Fir	Lodgepole Pine	PonderosaPine	Cottonwood/Willow	Aspen	Douglas-fir	Krummholz
Cache la poudre	0	3026	21454	2747	0	9741	0
Comanche Peak	87528	125093	14300	0	5712	7626	13105
Neota	18595	8985	0	0	0	0	2304
Rawah	105717	146197	0	0	3781	0	5101

FIGURE 17 – Table de contingence zone forestière/ Espèces

Les quatre comparaisons précédentes sont celles que nous avons jugées les plus pertinentes du fait qu'elles discriminent le plus les quatre zones. Nous allons donc caractériser nos zones à partir de l'altitude, de la distance horizontale à une route, de la distance horizontale à un départ de feu et de l'espèce d'arbre présente. Nous pouvons dresser la taxonomie suivante :

- "Cache la poudre" est une zone basse, influencée par la présence humaine puisqu'elle est homogènement proche des routes et des départs de feu. Elle présente une couverture forestière singulière puisqu'elle ne comprend ni Spurce/fir, ni Aspen, ni Krummholtz, mais en revanche riche en pin Ponderosa, Cottonwood et Douglas/Fir.
- "Comanche Peak" est une zone haute et plutôt sauvage qui se caractérise par la forte diversité des espèces présentes sur son territoire puisque seuls les Cottonwood en sont absents.
- "Neota" est une zone très homogènement haute, paradoxalement peu sauvage et qui a peu brûlé. Elle ne présente qu'une faible palette d'essences sur son territoire puisqu'elle comporte essentiellement des Spurce et des pins Lodgepole. Il est également notable que c'est la moins observée des quatre zones.
- "Rawah" est une zone haute, très sauvage, dont la couverture forestière se caractérise par la forte présence de Spurce et de pins Lodgepole, ainsi que la présence d'Aspen.

Nous avons donc deux zones particulièrement semblables qui sont Commanche Peak et Rawah. Ces deux zones ont été les plus étudiées, probablement parce qu'elles sont plus sauvages et donc plus proches des forêts primaires. Par ailleurs, nous avons une zone

particulièrement atypique en la présence de Cache la poudre, ce qui s'explique par le fait qu'elle est une zone nettement plus basse que les autres.

Influences des autres variables sur la couverture forestière

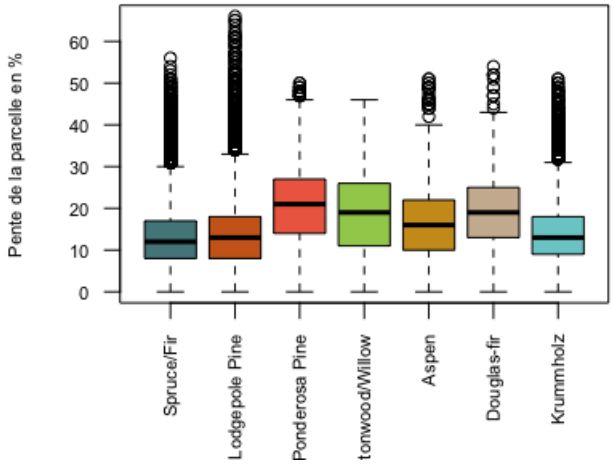


FIGURE 18 – Degré d'inclinaison du terrain en fonction du type d'espèce

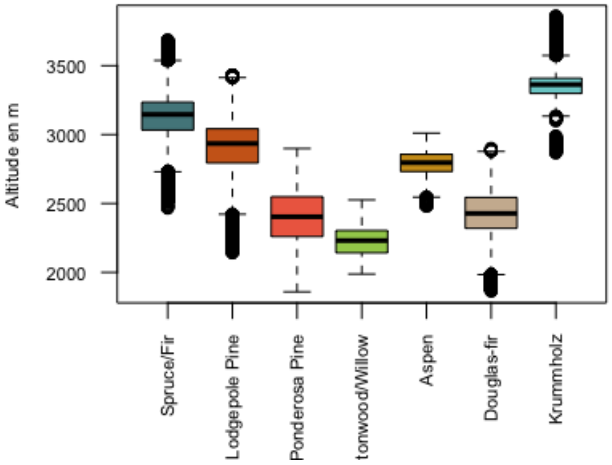


FIGURE 19 – Altitude en fonction du type d'espèce

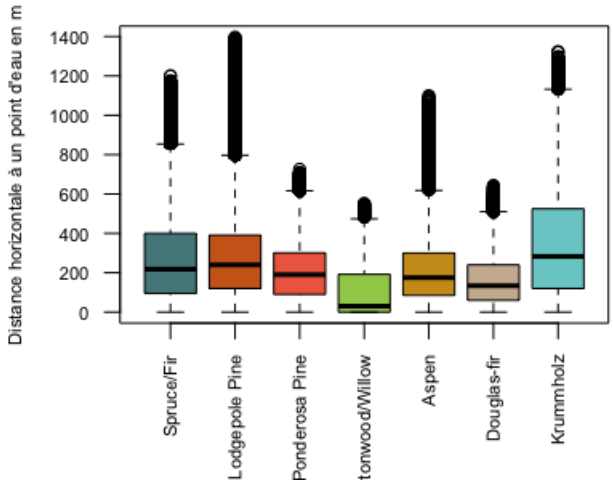


FIGURE 20 – Distance horizontale à un point d'eau fonction du type d'espèce

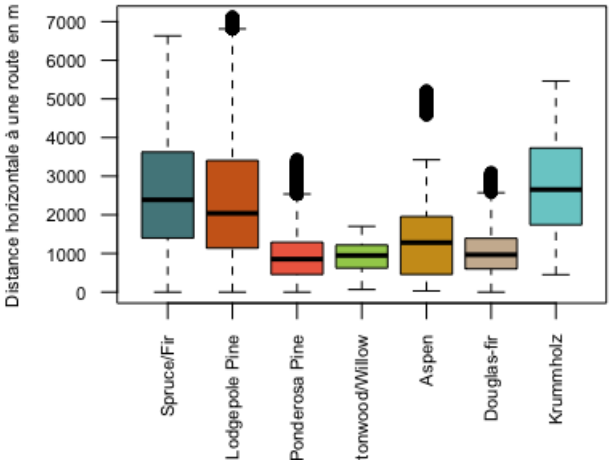


FIGURE 21 – Distance horizontale à une route en fonction du type d'espèce

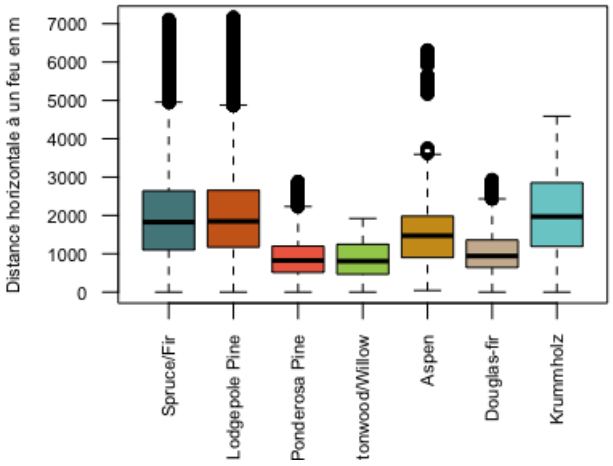


FIGURE 22 – Distance à un point départ de feu en fonction du type d'espèce

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Spruce/Fir	0	0	0	182	0	0	0	43	161	956	747	2693	2197	0	0	636	214	70	2461	3717
Lodgepole Pine	0	852	1191	3251	0	912	105	136	986	10803	9077	27278	13258	0	0	1743	957	1659	1490	5207
Ponderosa Pine	2101	4991	2411	7501	967	3993	0	0	0	11532	1353	0	41	116	0	129	506	0	0	2
Cottonwood/Willow	178	115	1018	168	48	320	0	0	0	224	34	0	0	155	0	51	436	0	0	0
Aspen	0	264	0	585	0	0	0	0	0	260	681	0	1315	0	0	35	600	170	67	53
Douglas-fir	752	1303	203	631	582	1350	0	0	0	8859	518	0	614	328	3	251	709	0	0	280
Krummholz	0	0	0	78	0	0	0	0	0	0	0	0	6	0	0	0	0	0	3	0
	21	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Spruce/Fir	804	25783	35557	11164	125	283	604	43	41911	7644	11863	21358	18148	94	931	14	0	8729	7882	4826
Lodgepole Pine	21	7442	20761	9702	349	2174	451	891	71399	20218	13209	29556	25308	1431	12	42	0	740	358	332
Ponderosa Pine	0	0	0	0	0	0	0	0	0	0	0	106	5	0	0	0	0	0	0	0
Cottonwood/Willow	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Aspen	0	0	699	70	0	132	0	12	1132	2111	309	460	518	20	0	0	0	0	0	0
Douglas-fir	0	0	29	138	0	0	0	0	0	0	63	200	539	15	0	0	0	0	0	0
Krummholz	13	148	706	204	0	0	31	0	805	197	222	839	636	51	948	63	298	6104	5566	3592

FIGURE 23 – Table de contingence Espèces/ Sol

Cette fois-ci, des douze comparaisons observées entre la variable "espèce" et les autres variables, nous avons choisi d'en retenir six. Nous avons donc mis en lumière les paramètres qui influencent la présence de telle essence plutôt qu'une autre. Parmi ceux-ci nous dénombrons donc la pente, l'altitude, la distance horizontale à un point d'eau, la distance à une route, la distance à un départ de feu et enfin le type de sol. Nous laissons ici de côté les liens avec les zones sauvages qui a déjà été exploré précédemment. Nous relevons donc ci-après, pour chaque essence d'arbre, les facteurs qui semblent influencer favorablement leur présence sur une parcelle.

- Les Spurce/Fir semblent se plaire à des altitudes hautes, dans des zones éloignées de l'activité humaine comme semble l'indiquer les figures 19 et 21, dont il est notable qu'elles sont fort semblables.
- Les Lodgepode Pine s'épanouisse dans des régions hautes également (environ 3000 mètres d'altitude en moyenne) et dans des zones sauvages (figures 21 et 22). Ils pousseront sur une grande diversité de sol.

- Les Ponderosa Pine apprécient les terrains pentus de moyenne altitude (environ 2300 mètres en moyenne), relativement proche de l'activité humaine (figures 21 et 22). Il ne pousse que sur un nombre restreint de sols (10 en tout).
- Les Cottonwood/Willow poussent sur des terrains comportant beaucoup de similitudes avec ceux accueillant les Ponderosa Pines. Terrains pentus, de moyenne altitude (environ 2000 m), plutôt proche de l'activité humaine et sur des sols quasi identiques à ceux sur lesquels le Ponderosa Pine pousse. Ce qui les distingue est le besoin d'humidité des Cottonwillow.
- L'aspen se retrouvera sur des parcelles pentues, d'altitude se concentrant fortement autour de 2800, 2900 mètres. C'est une espèce qui s'adapte à la présence humaine comme à son absence et ne peut pousser que sur un nombre restreint de types de sols.
- Le Douglas/Fir peut s'observer sur des zones pentues de "basse altitude" relativement à notre étude (environ 2400 m en moyenne). Il s'adapte bien à l'influence humaine et peu aux variations de sols. Il est notable qu'il est le seul à pousser
- Le Krummholz, enfin, va se trouver sur des parcelles de très haute altitude (près de 3500 mètres en moyenne) et éloigné de l'activité humaine. Il se montre adapté à diverses condition d'humidité. Les sols qui le verront pousser sont majoritairement de type Leighcan.

Parmi nos six critères, nous pouvons mettre en avant deux critères particulièrement discriminants : l'altitude et le type de sol. Ils permettent premièrement quasiment à eux seuls de différencier les parcelles en fonction de leur couverture forestière. On peut en outre partitionner nos espèces en trois grands groupes :

1. les espèces de haute altitude, appréciant les terrains moins pentus, moins humides et plus sauvages, poussant sur des sols présentant des types correspondant à la seconde moitié de notre index : le Spurce/Fir, le Lodgepole Pine et le Krummholz.
2. les espèces de basse altitude, très pentue, plus humides et plus tolérantes à l'influence humaine, poussant sur des sols correspondants à la première partie de notre index : Le Panderose Pine, le CottonWood et le Douglas-Fir.
3. une espèce aux caractéristiques intermédiaires par rapport à celles qui viennent d'être évoquées mais qui font de l'Aspen une espèce singulière.

Caractérisation des types de sol

Les boîtes à moustaches comparées de l'altitude, pente, distance horizontale à un point d'eau, à une route, à un départ de feu et ainsi que les différentes variables caractérisant la luminosité en fonction du type de sol mettent en lumière une relation entre ces variables. Ceci est logique puisque des types de sol sont plus favorables à une zone géographique montagneuse et humide ou alors à une zone non montagneuse et sèche. La majorité des variables ont donc une influence sur le type de sol.

Prolongement

A l'instar des deux chercheurs ayant travaillé sur le jeu de données nous pensons qu'un éventuel futur travail de modélisation devrait se proposer de classifier les parcelles en

fonction de leur type de couverture forestière à savoir des espèces qui y poussent. La présente analyse à permis de sélectionner deux variables particulièrement discriminantes (l'altitude et le type de sol) ainsi que quatre secondaires (la pente, la distance horizontale à un point d'eau et la distance horizontal à un feu). Ceci dit, en fonction du modèle utilisé et de notre puissance de calcul, il pourrait se montrer fructueux de tenir compte des 12 variables explicatives.