Compte rendu final du projet de SY09 sur le Bigfoot

Aurélie Law-Yen - Laura Miguel

juin 2023

Résumé

Le but du projet de l'UV SY09 (Science des données) est d'appliquer les méthodes étudiées en cours sur un jeu de données du répertoire Github Tidy Tuesday. Les données choisies [4] sont des témoignages de personnes qui auraient relevé la présence du Bigfoot aux États-Unis. Elles proviennent du site de la Bigfoot Field Researchers Organization (BFRO). Ce rapport expose les enjeux de l'étude, l'exploration et l'analyse des données.

1 Introduction

Le Bigfoot est une créature large, poilue qui vivrait dans les régions sauvages des États-Unis. Depuis 1995, il est étudié par le Bigfoot Field Researchers Organization (BFRO).

Les données étudiées sont tous les témoignages antérieurs à 2022 et que la BFRO a vérifié par des entretiens avec les témoins. Les témoignages sont constitués d'un rapport textuel, d'informations sur la localisation, la saison, la date, ainsi que de données météorologiques. Par ailleurs, une classification de ces témoignages a été établie :

A : Le témoin a vu le Bigfoot.
B : Le témoin a croisé le Bigfoot mais ne l'a pas distinctement vu.
C : Le témoignage manque de précisions.

Le système de classification des rapports de la BFRO évalue s'il y a potentiellement eu une erreur d'interprétation sur les signes de présence du Bigfoot. Cela amène donc à orienter ce projet sur l'étude de cette classification : au vu du contexte et du témoignage dans le rapport, peut-on retrouver si le témoignage appartient à la classe A ou B ? Si c'est le cas, on cherchera quelle(s) variable(s) influencent cette classification. Cela permettrait d'une part de classifier automatiquement les futurs témoignages, et de l'autre, de connaître les facteurs permettant de savoir avec certitude si l'on a croisé Bigfoot.

Ce rapport présente dans une première partie la transformation et l'exploration du jeu de données, puis une analyse des données d'abord grâce à des méthodes d'apprentissage non-supervisé, puis des méthodes d'apprentissage supervisé, et enfin à une analyse du texte. Finalement, une discussion sur les résultats obtenus sera apportée.

2 Exploration des données

2.1 Transformation des données

Le jeu de données contient 5021 témoignages auxquels sont associés 28 variables. Des variables ont été ajoutées comme le jour, le mois et l'année à partir de la variable date et les coordonnées polaires r et *theta* à partir de la *longitude* et de la *latitude* (utile si les classes sont liées à l'éloignement à un centre [1]). Il est nécessaire de convertir certaines variables vers des unités utilisées en France pour faciliter l'exploration et les interprétations des données (températures passant du Fahrenheit au degré Celsius, miles transformés en kilomètres, etc.). Pour maximiser le potentiel des données utilisables, les variables catégorielles liées aux saisons et aux types de précipitations ont été transformées en variables numériques par codage disjonctif complet.

Les données sont incomplètes pour un certain nombre de témoignages (seulement 721 témoignages avec aucune variable nulle). Les valeurs manquantes sont remplacées par les valeurs moyennes. L'un des cas particuliers concerne la *longitude* et la *latitude* remplacées par la valeur moyenne par État, et le jour, le mois et l'année, prenant les valeurs médianes (pour avoir des valeurs discrètes) par saison. Cela permet de garder plus de cohérence dans les valeurs remplacées.

Le texte ne sera considéré que dans un second temps.

2.2 Analyse des données

Les variables prises de façon indépendante correspondent aux conditions météorologiques moyennes des États-Unis. Le Bigfoot est particulièrement croisé en été et en automne, l'hypothèse est que la météo y est

favorable au sorties dans les espaces naturels et que des évènements folkloriques tel que Halloween y ont lieu. Au niveau de la localisation (figure 1), les États comptant le plus de témoignages sont Washington, Californie, Ohio et Floride et semblent correspondre à des zones montagneuses et/ou marécageuses.

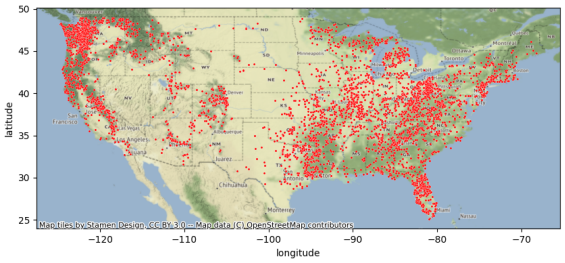

FIGURE 1 – Carte des États-Unis : répartition des rencontres avec Bigfoot.

Les plus vieux témoignages remontent à 1869, mais ils sont de plus en plus fréquents à partir des années 1960 et en particulier à partir de 1967. Cela peut s'expliquer par la sortie du film (présenté comme un documentaire) *Bigfoot* [3]. Dans les années 90, il y eut un net engouement (1995 étant l'année de création de la BFRO) jusqu'à 2004, l'année avec le plus de témoignages.

La proportion des témoignages dans la classe A et la classe B est sensiblement la même (table 1). Nous avons fait le choix de ne pas prendre en compte les témoignages de la classe C étant donné le faible nombre de rapports pour cette classe. De façon inattendue, même la classe B est associée à de très bonnes visibilités (environ 13.7 km).

TABLE 1 – Nombre de témoignages par classe

Classe	Nombre
classe A	2481
classe B	2510
classe C	30

2.3 Matrice de corrélation

Une matrice de corrélation a été calculée (figure 2) et la variable *class_A_B* a été spécialement créée et prend pour valeur 1 lorsque la classe de *classification* vaut *A* et 0 lorsqu'elle vaut *B*. La carte de chaleur ne permet pas de mettre en évidence une relation entre la *classification* et une autre variable prise individuellement. En effet, en valeurs absolues, les corrélations liées aux classes A et B et sont faibles. Il ne semble donc pas pertinent de sélectionner un sous ensemble de variables à étudier à partir de cette visualisation. Cependant, on peut voir que d'autres variables sont fortement corrélées entre-elles, en particulier *temperature_high* et *temperature_low* sont fortement corrélées à la température moyenne, et la *longitude* est corrélée avec la variable *r*. De plus, *geohash* est redondant avec la *latitude* et *longitude*, et le point de rosée *dewpoint* dépend de la température et de la pression. On peut donc supprimer ces variables.

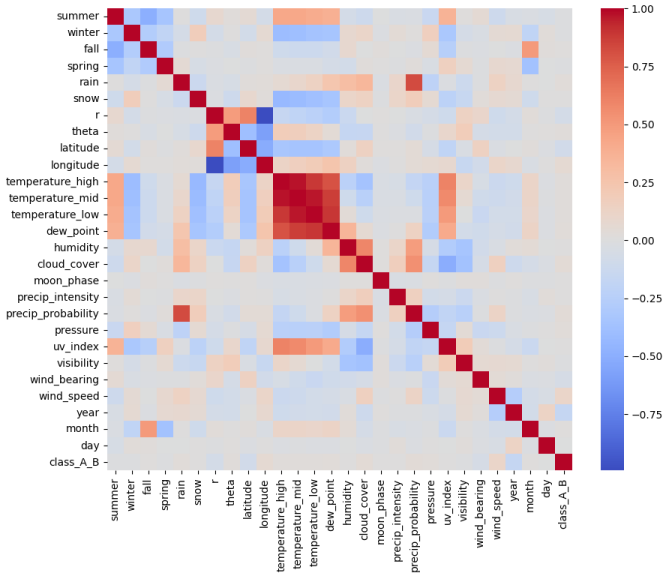

FIGURE 2 – Carte de chaleur des corrélations

3 Méthodes non-supervisées

3.1 Analyse Factorielle de Données Mixtes (AFDM) [2]

Dans un premier temps, nous avons cherché à représenter nos données dans une dimension réduite, voir si des variables sont corrélées, et si une projection 2D permet d'observer une frontière entre les classes. Nos données étant à la fois des données numériques (par exemple la température) et catégorielles (par exemple les différentes saisons), nous avons dû appliquer une AFMD : il s'agit d'une ACP avec des changements préalables sur les données. En particulier, les valeurs numériques ont été centrée-réduites pour que toutes les variables aient la même importance , et les variables catégorielles (après codage disjonctif complet) sont remplacée par la racine de la probabilité de tomber dessus (le but est de

ramener la variance à 1). C'est sur ce jeu de données transformé qu'on appliquera l'ensemble des méthodes. Une fois ces changements faits, on peut lancer une ACP sur les données, qui va permettre de réduire la dimension en minimisant les écarts entre les points originaux et leur projection sur les nouveaux axes. Les résultats montrent que les deux premières composantes principales n'expliquent que 23% de la variance : en essayant de représenter les données dans le premier plan factoriel, nous ne retrouvons pas spatialement les classes A et B.

3.2 Classification Hiérarchique Ascendante (CAH) et méthode des centres mobiles

Nous avons ensuite déroulé une classification hiérarchique ascendante avec le critère de Ward : en partant d'un ensemble où il y a autant de classes que d'individus, l'algorithme fusionne successivement des classes de façon à minimiser l'augmentation d'inertie intra-classe. Le dendrogramme résultant a montré un grand saut entre 8 et 9 classes, ce qui laisse penser que nos données peuvent être correctement séparées en 8 classes, or nous n'en cherchons que 2. Nous avons ensuite utilisé la méthode des centres mobiles en cherchant à identifier 2 classes. L'algorithme part de 2 points tirés au hasard, associe les autres points au point initial dont ils sont le plus proches, calcule le centre de gravité de la classe ainsi construite, et recommence avec comme points de départ les centres jusqu'à ce que ça ne bouge plus. Il cherche à minimiser l'inertie intra-classes à chaque itération (minimisation locale). Le résultat dépendant des centres de départ, on le fait tourner 10 fois et on garde le meilleur résultat en terme d'inertie. Graphiquement, on ne retrouve pas les classes A et B (figure 3). Comme la CAH laissait penser à la présence de 8 classes, on a essayé avec 8 classes en imaginant que des classes différentes représentaient peut-être toutes les deux la classes A ou B : ce n'est pas le cas.

À ce stade de notre analyse, nous avons découvert que les données météorologiques n'ont pas été récupérées sur le site de la BFRO, mais à l'aide d'une API à partir de la date du témoignage et de sa localisation. De plus, les témoignages sont laissés sur le site de la BFRO, et, même s'ils sont vérifiés un par un, ils ne relèvent pas d'une démarche rigoureuse comme en atteste les nombreuses données manquantes. Pour la suite de l'analyse, nous ne nous attendons pas à retrouver les 2 classes, mais nous avons tout de même essayé les différentes méthodes abordées en cours.

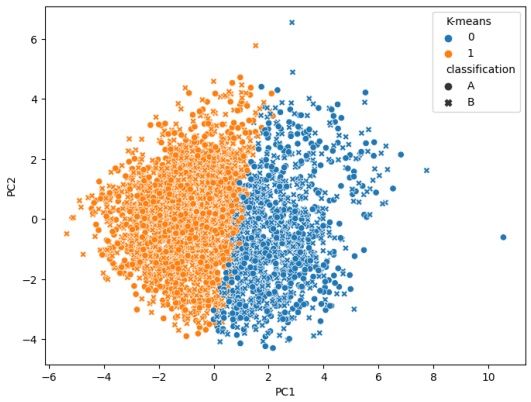

FIGURE 3 – ACP et K-means : résultat du K-means sur les données représentées dans le premier plan factoriel

4 Méthodes supervisées

4.1 K plus proches voisins

La méthode des K plus proches voisins permet de classifier de nouveaux individus en les associant à la classe la plus représentée parmi les K plus proches individus dont on connaît la classe voisins. Pour connaître le K à choisir, on effectue la méthode sur un intervalle de voisins (1 à 500), et on garde le nombre de voisins où le moins d'erreurs sont commises. On choisit de réaliser une validation croisée à 10 plis qui permet d'estimer les performances d'un modèle en utilisant l'ensemble des données disponibles, avec des données séparées en 10 sous-ensembles, avec donc un dixième servant d'ensemble d'évaluation et changeant à chaque itération de la validation croisée. En effet, notre jeu de données n'est pas suffisamment grand pour créer 3 ensembles entraînement, validation et test. Pour évaluer les performances des modèles, on utilise la précision (nombre prédictions correctes sur le nombre de prédictions totales). La meilleure trouvée est 0.55 pour 152 voisins : cela revient au même que d'attribuer une classe au hasard, donc ce modèle ce révèle inutile.

4.2 ADL, ADQ, bayésien naïf, régression logistique

Dans l'objectif d'appliquer les méthodes vues en cours malgré les résultats obtenus précédemment, une validation croisée à 10 plis a été appliquée sur une analyse discriminante quadratique (ADQ), une analyse discriminante linéaire (ADL) et un classifieur bayésien naïf. Ce sont trois méthodes qui font des hypothèses sur la distribution des données (notamment, suivi d'une loi nor-

male). Pour les trois méthodes, la précision est proche de 0.5 : c'est toujours équivalent à du hasard.

La régression logistique permet de séparer deux ensembles en utilisant une fonction logistique pour modéliser la relation entre les variables d'entrée et la probabilité d'appartenir à une classe donnée. La précision obtenue (0.56) n'est pas meilleure que précédemment.

4.3 Méthodes arborescentes

Dans le cadre des méthodes arborescentes, les arbres de décision et une méthode de *bagging* basée sur les arbres de décision sont testées. Les arbres de décision divisent le jeu de données en considérant les variables une par une et en choisissant la meilleure séparation selon un critère d'impureté. Le *bagging* (venant de "Boostrap Aggregating") basé sur les arbres de décision est une technique combinant plusieurs arbres de décision entraînés sur des échantillons indépendants du jeu de données pour obtenir des prédictions plus robustes. Par validation croisée sur 10 plis, on obtient que la précision de l'arbre de décision est de 0.51 et celle du *bagging* est de 0.49.

Table 2 – Précisions des méthodes supervisée sur les variables numériques et catégorielles

Modèle	Précision
K plus proches voisins	0.55
ADL	0.58
ADQ	0.54
Bayésien naïf	0.55
Régression logistique	0.56
Arbre binaire	0.52
Bagging	0.49

5 Analyse textuelle

Suite à ces tentatives de classification basées sur des valeurs numériques et catégorielles, nous nous sommes intéressées aux variables textuelles du jeu de données, à savoir dans un premier temps le titre du témoignage, puis le témoignage de la personne se présentant comme ayant croisé le Bigfoot (pour les deux variables, les mêmes méthodes ont été appliquées, seuls les résultats changent).

On commence par présenter les résultats sur les titres. Tout d'abord des tokens ont été extraits (on utilise le stemming pour récupérer des tokens, à savoir les mots du texte privés de leur suffixe). Cette étape nous a aussi permis de nous débarrasser des *stopwords* de langue anglaise : aussi appelés en français *mots vides*, ce sont des mots qui sont sans intérêt puisqu'ils sont toujours présents comme "is", ou "are". Ensuite, il a fallu sélectionner un ensemble de tokens à utiliser, l'idée étant de rajouter une colonne dans notre jeu de données pour chacun des mots sélectionnés. Enfin, deux approches différentes ont été testées pour remplir les colonnes : mettre 0 si le token est absent du texte, 1 sinon ; et remplir avec le TF-IDF du mot. L'analyse textuelle n'est pas couplée avec les variables numériques et catégorielles de base car lors de divers essais cela n'apportait pas de changement significatifs, sûrement car l'importance du texte dans ce problème de classification surpasse celle des autres variables (ce qui semble cohérent avec les résultats trouvés en se servant uniquement des variables numériques et catégorielles).

5.1 Sélection des tokens à utiliser comme colonnes

Dans un premier temps, on a compté pour chaque token combien de fois il apparaît dans la classe A et dans la classe B. D'après l'histogramme des mots les plus fréquents par classe, les dix premiers mots sont les plus représentatifs (table 3). À noter que pour la classe A le mot "sighting" et pour la classe B le mot "possible" sont bien plus présents que les autres ce qui est cohérent avec la signification des classes : pour la classe A, le témoin a vu Bigfoot, et la classe B, il ne l'a pas vu distinctement mais l'a possiblement croisé.

Table 3 – 10 mots les plus présents au sein du corpus par classe pour la variable Titre

Mot	Occurrence
sighting	904
encounter	256
daylight	254
creature	240
road	230
night	186
motorist	178
large	157
river	148
nighttime	146

Table 4 – classe A

Mot	Occurrence
possible	849
vocalizations	325
sighting	264
heard	233
lake	222
outside	184
hear	182
night	179
hears	171
encounter	165

Table 5 – classe B

Ces dix mots les plus présents pour chaque classe sont utilisés, on obtient donc 17 variables car des mots sont communs aux deux listes de mots établies.

5.2 Codage disjonctif complet

Ces 17 variables sont codées avec un encodage disjonctif complet, puis réduites pour réaliser une ACP. Même si les deux premiers axes factoriels obtenus ne représentent que 20% de la variance expliquée, le cercle des corrélations associé au premier plan factoriel (figure 4) permet de visualiser dans le quart en haut à gauche les mots les plus représentatifs de la classe A selon la tableau d'occurrence et dans le quart en haut à droite ceux associés à la classe B. Cela permet de supposer que ces 17 variables permettent de classifier les données en deux classes de façon plus satisfaisante qu'avec les données numériques et catégorielles.

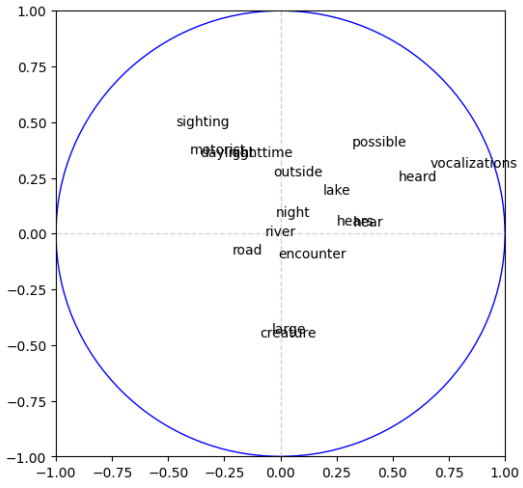

FIGURE 4 – Cercle de corrélation du premier plan factoriel sur l'ensemble des mots les plus présents dans les classes A et B pour la variable Titre

À partir de ces 17 variables, des validations croisées à 20 plis ont été appliquées sur diverses méthodes dans le but d'effectuer une classification binaire : ADL, ADQ, bayésien naïf, régression logistique, arbre de décision et forêt aléatoire. Ces précisions sont supérieures à toutes celles obtenues dans les parties précédentes où on n'utilisait pas le texte, et les meilleures précisions ont été obtenues (figure 5) avec la régression logistique avec un score de 0.765 et pour un écart-type plus faible que pour les autres méthodes.

FIGURE 5 – Précisions pour diverses méthodes de classification appliquées sur notre sélection de tokens de la variable Titre

On se penche donc sur le classifieur lié à la régression logistique, où il est possible d'interpréter le résultat en analysant l'impact des différentes variables sur la probabilité de la classe positive. En effet, les coefficients associés à chaque variable composant β sont accessibles (table 6), ce qui permet d'interpréter leur effet relatif sur la prédiction de la classe. Plus le coefficient est inférieur à 0, plus il a de poids pour classer le témoignage en classe A, et inversement, plus il est supérieur à 0, plus il a de poids pour classer le témoignage en classe B. Les coefficients obtenus sont cohérents avec les mots les plus présents par classe. Cependant, par classe, ce ne sont pas les mots avec le plus d'occurrences qui ont les coefficients les plus importants en valeur absolue. Par exemple, *hears* est le 9ème mot le plus présent des titres de la classe B mais est le 3ème mot donnant le plus de poids pour classifier un tire dans la classe B avec la régression logistique proposée. Cela pourrait s'expliquer par l'absence de ces mots ou de leur faible présence dans les témoignages de l'autre classe. Les mots dont le coefficient est proche de 0 semblent être les tokens qui apparaissent aussi souvent dans une classe que l'autre, par exemple *night*) : c'est un résultat auquel on s'attendait.

On peut donc retenir comme méthode de classification des témoignages du Bigfoot entre les classes A et B, le classifieur de régression logistique avec pour variables, en codage disjonctif complet, les 17 tokens obtenus par regroupement des 10 tokens les plus présents par classe.

5.3 TF-IDF

On a également essayé une autre façon de donner les valeurs pour les colonnes formées pour avec les tokens

TABLE 6 – Coefficients β de la régression logistique effectuée classés par ordre décroisant et associé à la variable Titre

mot	coefficient		
sighting	-6.22	lake	0.25
creature	-2.71	outside	0.41
road	-2.59	large	0.44
encounter	-1.89	vocalizations	2.91
daylight	-1.54	heard	3.03
river	-0.92	hears	3.80
nighttime	-0.72	hear	4.12
motorist	-0.41	possible	10.82
night	0.05		

avec la métrique *TF-IDF* (*Term Frequency - Inverse Document Frequency*) : c'est une mesure statistique de l'importance d'un mot dans un document relativement à un ensemble de documents qu'on peut appeler un corpus. Le *TF* permet d'apprécier à quel point le terme est présent au sein d'un témoignage et le *IDF* exprime l'importance du terme par rapport à l'ensemble du corpus. Plus l'*IDF* prend une grande valeur, plus le mot est spécifique au témoignage. Donc *TF-TDF* (multiplication entre *TF* et *IDF*) représente la pertinence d'un témoignage par rapport à un token, en particulier parce que l'*IDF* pondère le fait qu'un token peut simplement être très présent dans tout le corpus et que dans ce cas la seule analyse de *TF* n'est pas suffisante. Les formules suivantes détaillent son expression de façon adaptée au sujet :

$$\text{TF} = \frac{\text{Occurrence du token dans le témoignage}}{\text{Nombre total de tokens dans le témoignage}}$$

$$\text{IDF} = \log\left(\frac{\text{Nb de témoignages}}{\text{Nb de témoignages contenant le token}}\right)$$

Sur la base de cette métrique, les mêmes méthodes sont appliquées que dans la partie précédente en remplaçant les 1 de chaque variable par la valeur de *TF-IDF* du mot. On retrouve des résultats similaires.

5.4 La variable contenant le témoignage textuel

En voyant les tokens ressortant avec le titre, on s'est douté que le titre était donné par la même personne donnant la classification : *possible* ne se retrouvait que dans les témoignages de la classe B. Nous avons donc refait la même méthodologie faite sur la variable *title* mais avec le témoignage textuel *observed*. Les tokens à utiliser s'en retrouvent donc changés en parti pour devenir moins évidents que *possible* ou *sighting* (par exemple, *creature*, *saw*, *heard*, *sound*), et nous avons eu des résultats légèrement moins bons (entre 0.65 et 0.71), ce qui paraît cohérent avec notre hypothèse sur la construction du titre.

6 Conclusion

Ainsi, le problème de classification des témoignages du Bigfoot dans les classes A ou B est possible en analysant les variables textuelles telles que le titre du témoignage ou le témoignage textuel. La régression logistique appliquée aux mots des titres les plus présents par classes a une précision de 0.765.

Au début du projet, suite à l'analyse exploratoire des données, on pensait que les données numériques ou catégorielles comme les données temporelles ou les données météorologiques seraient suffisantes. Hors ce n'est pas le cas, donc chercher à définir le meilleur moment et endroit pour voir le Bigfoot n'est pas possible avec des données comme le relevé de la date, des conditions météorologiques, etc. Il semblerait que cette classification en classe A et B soit subjective et dépende de la perception humaine du phénomène, ce qui expliquerait pourquoi le texte prévaut.

L'une des pistes pour approfondir l'analyse du texte serait de chercher si d'autres moyens de sélection des tokens donnent de meilleurs résultats. Toujours dans le but de déterminer les conditions propices pour voir le Bigfoot, on pourrait utiliser des réseaux de neurones, des techniques basées sur les transformers prenant en compte le contexte de suites de mots, de phrases, et non pas de simples analyses statistiques sur le texte. De plus, nous n'avons pas étudié la variable contenant une description textuelle du lieu : elle pourrait s'avérer utile.

Références

[1] Feature engineering for machine learning : 10 examples.

[2] The ultimate guide for clustering mixed data.

[3] R. Gimlin and R. Patterson. Bigfoot, 1967.

[4] T. Renner. Bigfoot sightings - dataset by timothyrenner.