Compte rendu intermédiaire du projet de SY09 sur le Bigfoot

Aurélie Law-Yen - Laura Miguel

28 avril 2023

Résumé

Le but du projet de l'UV SY09 (analyse de données) est d'appliquer les méthodes étudiées en cours sur un jeu de données du répertoire Github Tidy Tuesday. Les données choisies [2] 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 intermédiaire expose les enjeux de l'étude, et présente une première étape d'exploration et analyse des données.

1 Introduction

Le Bigfoot est une créature légendaire, large, poilue, mi-homme, mi-animal, qui vivrait dans les régions sauvages des États-Unis. Depuis 1995, il est étudié par le Bigfoot Field Researchers Organization (BFRO) se définissant comme la seule organisation scientifique consacrée à l'étude du Bigfoot.

Les données étudiées sont tous les témoignages antérieurs à 2022 et sur lesquelles la BFRO a mené un minimum d'investigations. 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 croisé le Bigfoot.
B : Il a possiblement croisé le Bigfoot.
C : Témoignage peu fiable.

Cela amène donc à orienter ce projet sur l'étude de la confiance pouvant être accordée à un témoignage. Au vu du contexte et des informations données par un témoin, 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. En connaissant ces facteurs, on pourra faire une hypothèse sur l'existence du Bigfoot.

Ce rapport intermédiaire présentera dans une première partie la transformation et l'exploration du jeu de données, puis une analyse en composantes principales, une classification ascendante hiérarchique et le résultat d'un K-means. Enfin, on apportera une discussion sur les résultats obtenus jusqu'à présent.

2 Transformation et exploration des données

Le jeu de données contient 5021 témoignages auxquels sont associées 28 variables. Les principales étapes de transformation ont été de supprimer des variables (geohash redondant avec la latitude et longitude), d'en ajouter (ajout du jour, mois et année à partir de la variable date) et d'en convertir 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.).

Les données sont relativement incomplètes pour un certain nombre de témoignages (seulement 721 témoignages avec aucune variable nulle). Les variables sont en grande partie corrélées entre elles par ce qu'elles représentent : en Floride, il fait en moyenne plus chaud que dans un autre État situé au nord du pays ; la température de point de rosée dépend physiquement de la température ambiante et du taux d'humidité, etc.

Après analyse des variables prises indépendamment, il en ressort que le Bigfoot a en général été repéré dans les conditions suivantes : températures entre 9,7°C et 21°C, absence de pluie ou pluie faible, la phase de la lune semble peu importer, la pression atmosphérique et le taux d'humidité sont dans les moyennes des États-Unis. Ces résultats ne donnent pas de pistes particulières sur l'environnement de vie de Bigfoot.

Au niveau de la localisation (figure 1), les États comptant le plus de témoignages sont Washington, Californie, Ohio et Floride. On peut observer que le Bigfoot est principalement aperçu dans les régions montagneuses.

Les plus vieux témoignages remontent à 1869, mais ils sont de plus en plus fréquents à partir des années 1960

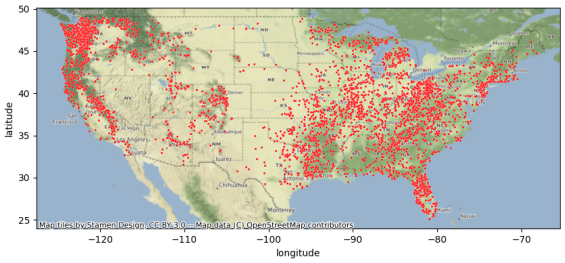

FIGURE 1 – Carte des États-Unis : répartition des rencontres avec Bigfoot.

et en particulier à partir de 1967. Cela peut s'expliquer par la sortie du film (présenté comme un documentaire) *Bigfoot* [1]. 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. De façon inattendue, même la classe B est associée à de très bonnes visibilités (environ 13.7 km). On note une différence de répartitions de témoignages en fonction des États, en particulier plus de classe B dans les États de Washington et Californie, et plus de classe A en Alabama, au Kentucky et au Texas. En suposant que Bigfoot vit effectivement dans les régions montagneuses, ce résultat est contre-intuitif puisque les États avec plus de classe A que B sont ceux qui sont principalement des plaines.

3 ACP, CAH et K-means

TABLE 1 – Variables pour l'ACP.

temperature_high	*dew_point*	*visibility*
temperature_mid	*moon_phase*	*pressure*
temperature_low	*cloud_cover*	*humidity*
precip_probability	*wind_speed*	*latitude*
precip_intensity	*wind_bearing*	*longitude*

Comme dit précédemment, les variables sont nombreuses et corrélées les unes avec les autres. Nous avons donc tenté de faire une ACP sur nos données quantitatives (table 1), avec pour objectif de représenter les données en 2 dimensions et d'essayer de retrouver les classes A et B. Avant de lancer l'ACP, les données ne variant pas sur les mêmes intervalles, il a fallu les normaliser et se débarrasser des lignes contenant des valeurs nulles (cela nous a laissé 2112 lignes). L'ACP nous a donné deux composantes principales expliquant respectivement 27% et 18% de la variance. En essayant de représenter les données dans le premier plan factoriel, nous n'avons pas pu retrouver spatialement les classes A et B.

On a déroulé une classification ascendante hiérarchique avec le critère de Ward et la distance euclidienne. Le dendrogramme résultant a montré un grand saut entre 1 et 2 classes ainsi qu'entre 2 et 3 classes. On a donc lancé un K-means avec 2 et 3 classes et regardé si les classes résultantes correspondaient aux classes A, B et C (figure 2) : ce n'est pas le cas.

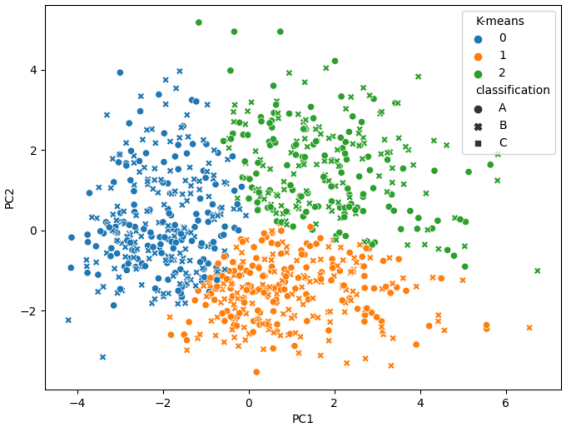

FIGURE 2 – ACP et K-means : résultat du K-means sur les données représentées dans le premier plan factoriel

4 Conclusion

Nous n'avons pas encore pu tirer de véritables résultats, mais cela nous a permis de réfléchir à la qualité des données. De fait, les données ne sont pas le résultat d'une démarche scientifique : les témoignages viennent de personnes individuelles ; même si la BFRO organise des expéditions, on a des données uniquement sur les fois où Bigfoot a été croisé (il aurait été intéressant de pouvoir comparer les expéditions réussies des expéditions ratées) ; etc.

Dans la suite de l'analyse, on essaiera d'appliquer des méthodes d'apprentissage supervisées (toujours dans l'objectif de retrouver les classes A et B), et on s'intéressera aux données textuelles.

Références

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

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