## Deuxième partie : Introduction

Dans la première partie de ce rapport nous avons explicité le contexte général des données sur lesquelles nous comptons effectuer nos analyses. En effet, nous y avons exposé en détails le jeu de données « *Forest Fires »*. Cette explication du contexte s’est conclue avec quelques questions ouvertes qui permettraient de mieux comprendre les feux de forêts qu’on peut observé dans le parc de Montesinho. Dans cette deuxième partie du rapport il sera question de soutirer de l’information de ces données afin de mieux comprendre ce qui caractérisent ces feux de forêts. Par la suite, nous utiliserons ces notions afin de répondre aux questionnements proposés dans la première partie du rapport. Afin de bien comprendre le phénomène à l’étude il est essentiel de débuter par une phase d’analyse statistique descriptive. En effet, les 13 variables qui composent notre jeu de donnée, ne sont pas intelligibles, c’est-à-dire qu’il est difficile de ressentir les tendances qui émanent de celle-ci. Autrement dit, il nous faut décrire les données en les résumant et en les structurant [4 p.192]. À partir de ces descriptions il nous sera possible de remarquer certaines propriétés qui caractérisent une forêt favorisant les incendies. Ensuite, nous exposerons les potentielles distributions probabilistes sous-jacentes à certaines de ces variables. Pour ce faire, nous utiliserons d’abord notre instinct. En effet, certaines de ces variables semblent d’emblée exprimée des signes de distribution normale. Nous utiliserons ensuite le test du chi carré afin de vérifier si il est probable ou non que ces variables soient effectivement issues d’une distribution normale. Ces vérifications seront fait sous la proposition d’une hypothèse nulle stipulant qu’elles suivent effectivement une loi normale. Comme nous le verrons, ces hypothèse s’avèreront improbable selon la *p-value* choisi. Afin de renforcir cette conclusion, nous utiliserons ensuite le teste de *Shapiro Wilk* selon la même p-value. Nous poursuivrons notre investigation concernant les distributions potentielles de certaines variables à l’aide d’une batterie de test un peu plus élaboré. En effet, afin de cibler certains candidats potentiels, nous analyserons un graph le graph de *Cullen et Frey* qui compare les propriétés d’aplatissement et d’asymétrie de certaines distributions avec les variables qui nous intéresse. À partir de ces résultats nous calculerons le critère d'information d'Akaike afin de trancher sur quel modèle continuer notre investigation. Enfin, nous passerons le test de Kolmogorov–Smirnov afin de rendre compte de la véracité probabliste de notre hypothèse sur les variables à l’édute. Nous verrons que certaines variables comme la température semble être exprimmer une distribution de type Weibull. Une fois que nous aurons cerner les distributions sous-jacentes à ces variables, ils nous sera possible d’estimer certains paramètres de la population d’où proviennent ces variables. Enfin, nous terminerons notre analyse statistique par une proposition d’un modèle de régression sur deux variables qui semblent partager une interaction. Finalement, nous conclurons en soulignant les résultats qui nous semblent pertinents et en prosant quelques quesitions ouvertes.

## 1) Analyse descriptive des données

Dans le but de bien cerner certaine tendance de l’état de la forêt lors d’incendie dans le parc de Montesinho, nous avons généré des histogrammes pour les variables qui composent le jeu de données. Nous avons joint à ces figures des légendes contenant certaines statistiques descriptives des variables à l’étude. Ces statistiques sont la moyenne, l’écart-type, le coefficient de symétrie et le coefficient d’aplatissement.

## 1.1) L’indice d’humidité de l’humus

Comme mentionné dans la première partie, le *DMC* représente l’indice d’humidité/facteur de combustion de la couche organique situé entre 5cm et 10cm sous le sol. Il est à noter que cette section du sol n’est pas affectée par les pluies de moins de 20mm. Cette section du sol est aussi protégée des périodes de sècheresse rapide. On remarque sans surprise que les instances de feux de forêts sur un sol ayant un indice de *DMC* très élevé se font plus rare. Par contre, on remarque une certaine tendance : le plus grand nombre d’instance de feux on une métrique de *DMC* qui semblent au milieu de l’échelle. On remarque aussi un écart type assez grand compte tenu de l’échelle en question. Les index d’humidité du humus sont donc plutôt dispersés dans un scénario de feux de forêts.

[4] W.W. Hines, D.C. Montgomery, D.M Goldsman, C. M. Borror, Probability and Statistic pour ingénieurs, 3ème éd. Montréal, Canada, Chanelière-Éducation, 2017.

[5] W.J. De Groot, Interpreting the Canadian Forest Fire Weather Iinde (FWI) System[EN LIGNE]. Disponible: http://www.dnr.state.mi.us/WWW/FMD/WEATHER/Reference/FWI\_Background.pdf