Etude de données - Animal Rescues in London

Branly Stéphane et Tran Quoc Hung

6 juin 2022

Résumé

Dans le cadre de l'UV SY09 (Science des données), un projet prend place dans le but d'appliquer les notions vues en cours sur un jeu de données. Le jeu de données sélectionné et celui de Animal Rescues in London proposé lors des challenges d'analyse de données #**TidyTuesday**. Ce document montre une analyse du jeu de données ainsi que le cheminement intellectuel que nous avons mené pour effectuer cette analyse. Cette analyse est guidée par des objectifs fixés. L'ensemble du code est disponible sur le repository Github SY09-Projet

1 Introduction

1.1 Présentation du jeu de données

Le jeu de données contient 31 variables et est disponible sous forme d'un fichier .csv. Ce jeu de données regroupe l'ensemble des interventions animalières (sauvetages d'animaux) comprenant des informations de localisation (code postal, arrondissement, quartier, coordonnées GPS), date et l'heure des interventions ainsi que des détails sur l'intervention (motif, lieu d'appel, animal, coût). Il s'agit des interventions enregistrées à Londres entre janvier 2009 et mai 2021.

1.2 Objectifs de l'étude

Étant donné le type de données : séries spatio-temporelles, nous avons décidé de comprendre l'implication des aspects spatial et temporel dans les interventions. Pour explorer les données, nous allons ainsi nous concentrer deux objectifs principaux.

1.2.1 Les animaux sauvés

Notre premier objectif sera de comprendre quels sont les animaux sauvés au cours des interventions. Et plus particulièrement de comprendre la relation du type d'animal en fonction de la localisation de l'intervention, de la date d'intervention, du coup ou encore du type d'intervention (*animal sauvé en hauteur, dans l'eau, ...*)

1.2.2 Les types d'intervention

Ce second objectif permettra de comprendre quels sont les types d'interventions et surtout comment ils dépendent de la localisation, de la date, ...

2 Exploration des données

2.1 Nettoyage des données

Avant de commencer à analyser les données, une première phase de nettoyage de données est nécessaire afin de faire en sorte qu'elles soient exploitables. Les 31 variables ont ainsi été analysées puis nettoyées. Le nettoyage a consisté à compléter les informations manquantes de localisation (*latitude, longitude*) grâce aux autres colonnes puis retirer des colonnes dupliquant l'information de localisation sous différentes formes. La matrice de corrélation nous a permis aussi de retirer des colonnes de relation linéaire triviale que nous n'allions pas utiliser pour le suite de l'étude (*hourly_notional_cost* et *pump_notional_cost*).

Pour la suite de l'étude, nous travaillons donc avec un jeu de données contenant 16 variables qui sont les suivantes :

TABLE 1 – Variables du jeu de données

date_time_of_call	animal_group_parent
pump_count	incident_notional_cost
originof_call	borough
property_type	property_category
special_service_type_category	special_service_type
latitude	longitude
month	year
dayofweek	hour

2.2 Analyse des données

2.3 Spatialité

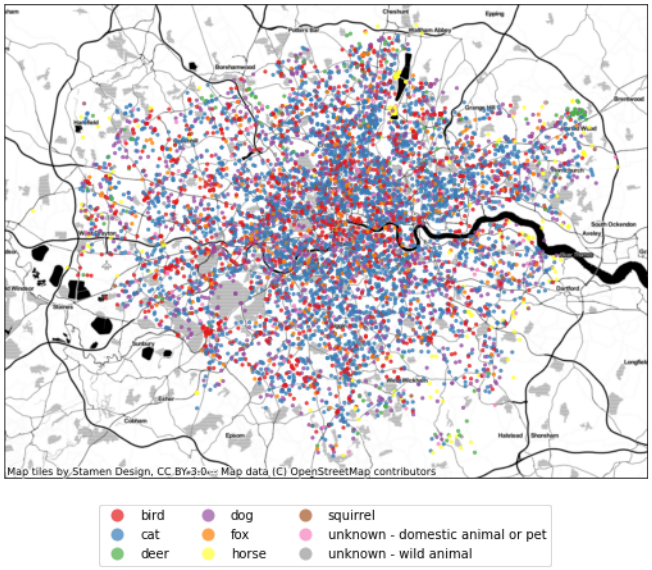

FIGURE 1 – Géolocalisation des interventions

Ensuite, nous nous sommes intéressés à la spatialité des interventions en fonction des animaux. A l'aide d'une carte (1 ainsi que celle disponible en annexes (9)), nous remarquons que les interventions d'animaux domestiques ou de petites tailles (*chats, chiens, oiseaux*) sont assez concentrées dans le centre de Londres. Tandis que les interventions sur des animaux ruraux / sauvages ou de grande taille (*chevaux, cerfs*) sont concentrées sur la périphérie de la ville.

2.4 Temporalité

Nous avons également décidé d'étudier la temporalité du jeu de données et en particulier voir les différences d'occurrences d'intervention en fonction des types d'animaux. Pour cela nous avons regardé à différentes échelles temporelles. Nous nous sommes concentrés sur les centres d'inerties par groupe d'animaux. Nous pouvons par exemple remarquer sur la figure 2 que les interventions d'écureuils (*squirrel*) ont principalement lieu en juillet et août tandis que les interventions sur des chiens sont uniformément distribuées sur les différents mois de l'année. Les analyses de centres d'inerties ont également étaient faites pour le jour de la semaine ainsi que pour l'heure.

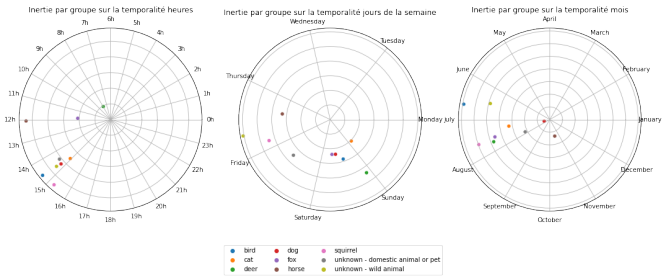

FIGURE 2 – Centre d'inertie temporel par type d'animal

2.5 Analyse en composantes principales

Après nettoyage des données, nous avons réalisé une analyse en composantes principales pour considérer la réduction le nombre de variables explicatives afin de permettre de visualiser les données vu leur haute dimensionalité. Nous remarquons que l'axe PCA_1 explique 27.6% de la variance totale et PCA_2 22.9%. De ce fait, avec ces deux composantes nous pouvons représenter 50.5% des informations contenues par les 16 variables. Il s'agit d'un faible pourcentage pour espérer effectuer de la compression de données. D'autant plus que l'allure des inerties cumulées est linéaire plutôt que logarithmique.

La visualisation des données sur les deux axes principaux de la PCA est visible sur la figure 3.

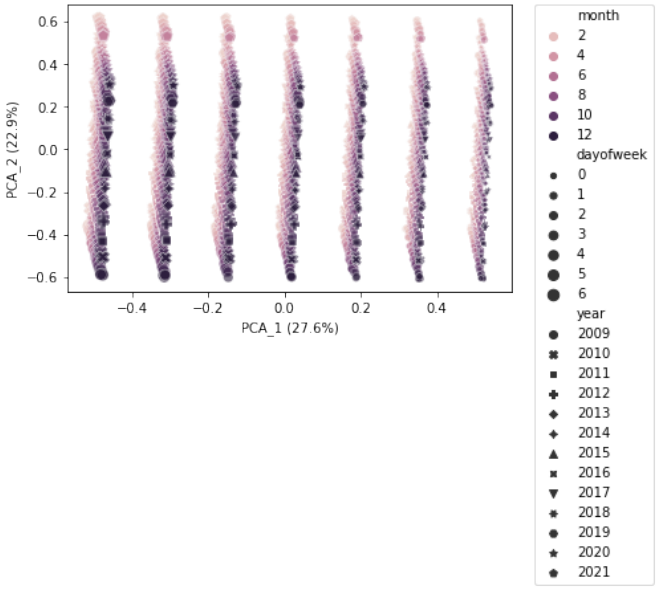

FIGURE 3 – Représentation du jeu de données sur les 2 principaux axes de PCA

Remarques : Nous avons ici des axes représentant principalement les variables temporelles. Il est intéressant de voir que la spatialité et coût d'intervention ne se distinguent pas. Idéalement, nous aurions voulu avoir une PCA permettant de distinguer l'animal ou encore le type d'intervention mais malheureusement nous ne distinguons pas de groupe pour ces variables qualitatives.

2.6 Partitionnement en k-moyennes

Nous avons également décidé de faire un partitionnement en k-moyennes sur notre jeu de données. Malheureusement le faible nombre de variables numériques fait que les classes créées ne nous permettent pas de les associer aux classes correspondantes aux *type d'intervention* ou *type d'animal*. Il en est de même pour notre jeu de données *amélioré* dont sa fabrication est détaillée dans la section 3.2.1, mais nous remarquerons l'impact de l'ajout de variables en rapport avec la spatialité. La visualisation du partitionnement en k-moyennes est visible en annexe 4.

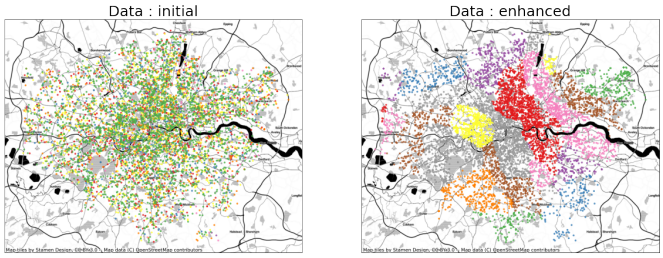

FIGURE 4 – Représentations des classes issues de la méthode de partitionnement par k-moyennes

2.7 Classification Ascendante Hiérarchique

Enfin, nous avons aussi une effectué une classification hiérarchique en groupant les jeux de données par *type d'animal* et par *type d'intervention*. Les données numériques utilisées correspondent aux moyennes. Il est intéressant de voir que le jeu de données *amélioré* rapproche les *chevaux, vaches, chèvres* ensembles ainsi que le *chat* du *renard*. Mais aussi que les interventions *en hauteur et au sol* sont éloignées des interventions *aquatiques*.

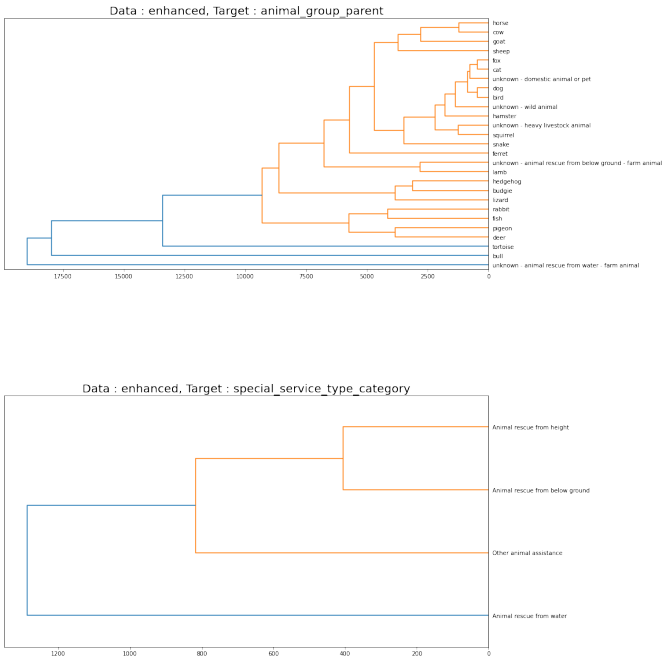

FIGURE 5 – Dendogrammes issus des classifications ascendantes hiérarchiques

3 Classification supervisée

3.1 Introduction

Après l'analyse des données, on présente des différents méthodes appliquées pour la résolution du problème de classification.

Ici, l'objectif de la classification est de prédire le **type d'intervention** puis le **type d'animal** que nous marquons en tant que Y. La variable à expliquer, y, est qualitative nominale à 4 modalités : *Animal rescue from water, Animal rescue from height, Animal rescue from below ground, Other animal assistance* pour le type d'intervention. Et à 10 modalités : *cat, bird, dog, fox, horse, unknown - domestic animal or pet, deer, unknown - wild animal, squirrel, unknown - heavy livestock animal* pour le type d'animal.

3.2 Préparation des données pour classification algorithmes

3.2.1 Apport d'une expertise

Nous avons pu appliquer certaines méthodes de classification supervisée mais les résultats n'étaient pas satisfaisants. Nous avons ainsi décidé **d'apporter une**

expertise dans le domaine notamment grâce à notre expérience, la lecture d'articles en rapport avec les interventions animalières à Londres ainsi que l'analyse du jeu de données. Nous avons ajouté des notions géographiques en indiquant pour chaque intervention la distance avec des objets physiques les plus proches (*foret, champs, lacs, zones commerciales, zones résidentielles, ...*). Cela permet d'ajouter des variables quantitatives qui ont un sens par rapport à nos cibles (*type d'animal et type d'intervention*). Les données géographiques utilisées sont issues de OpenStreetMap et permettent de récupérer des *features* par type de géométrie (*Point, Ligne, Multiligne, Polygone...* et par *tags*.

Les données ont été récupérées pour la ville de Londres (représentation sur la figure 6).

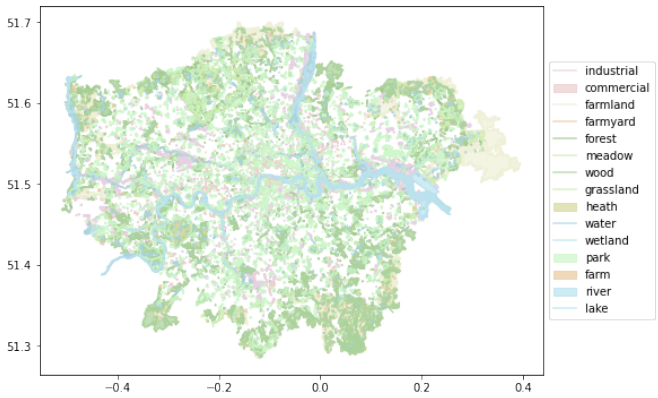

FIGURE 6 – Représentation des données géographiques récupérées pour Londres

Nous avons aussi décidé d'effectuer des transformations sur les variables temporaires. En effet l'ordre du *mois*, *jour de la semaine* ou encore *l'heure* faisaient que par exemple la distance de *janvier à décembre*, du *lundi à dimanche* ou encore de *00h à 23h* étaient plus grandes que les distances respectivement de *janvier à février*, du *lundi à mardi* ou encore de *00h à 01h*. Ces variables ont alors étaient transformées sous forme *cyclique sur 2 variables* afin de ne pas avoir de problèmes d'ordre. La figure 2 illustre la transformation effectuée.

Le jeu de données utilisant cette expertise et transformation sera nommé jeu de données **amélioré (enhanced)**, le jeu de données initial sera nommé **initial (initial)**.

3.2.2 Pré-traitement des données

La première étape de tout projet d'apprentissage automatique consiste à transformer les données dans un format pouvant être utilisé par des algorithmes d'apprentissage automatique. Dans la plupart des cas, on a besoin de données numériques sans valeurs manquantes ni valeurs aberrantes qui pourraient rendre beaucoup plus difficile l'apprentissage d'un modèle.

Dans le cas de l'ensemble de données, les valeurs sont déjà très nettoyées, mais on devra transformer les données en valeurs numériques.

3.2.3 One-Hot Encoding

Pour les variables qualitatives, nous avons effectués des transformations utilisant le One-Hot Encoding. Nous créons ainsi autant de variables booléennes que de modalités pour notre variable qualitative. La variable est mise à *True* s'il s'agit de la modalité, à *False* sinon.

3.2.4 Réduction de dimension, sélection de variables

Nous avons également décidé d'ajouter un jeu de données de dimension réduite en utilisant une méthode de sélection de variables. Nous avons utilisé notre jeu de données amélioré qui contient le plus de colonnes. La méthode de sélection de variables utilisée et celle d'élimination de variable récursive qui permet ensuite d'établir un rang d'importance pour chaque variable. Ce rang d'importance est calculé à l'aide d'un estimateur. Ici nous avons utilisé l'estimateur de régression logistique. La sélection de variables permet de diviser par 2 le nombre de variables. Nous verrons par la suite si cela altère les résultats des algorithmes.

Le jeu de données contenant les variables sélectionnées sera appelé **variables sélectionnées (selected features)**.

3.2.5 Partition des jeux de données

Nos jeux de données ont été partitionnés afin de pouvoir trouver les hyperparamètres, entraîner et comparer les méthodes avec de la cross-validation ainsi que pour tester les résultats. Ainsi nous avons mélangé les jeux de données (changement de place des individus) afin de ne pas conserver l'ordre temporel, puis nous avons séparé les jeux de données en trois parties. Nous prenons soin de garder la même graine aléatoire pour mélanger les individus issus de différents jeux de données (*initial, amélioré et variables sélectionnées*.

3.3 Méthodes utilisées

Nous avons utilisé plusieurs méthodes de classification supervisée pour nos jeux de données et targets.

3.3.1 Les hyperparamètres

Certaines de ces méthodes nécessitent de trouver des hyperparamètres comme la méthode des K plus proches voisins ou encore Random Forest.

Par exemple pour l'algorithme des K plus proches voisins, pour déterminer le nombre K optimal de voisins, nous avons utilisé une **grille de recherche (Grid Search)**, c'est une méthode d'optimisation (hyperparameter optimization) qui va nous permettre de tester une série de paramètres et de comparer les performances pour en déduire le meilleur paramétrage. Il existe plusieurs manières de tester les paramètres d'un modèle et le Grid Search est une des méthodes les plus simples. Nous vérifions ensuite que les K trouvés sont cohérents par rapport à un autre échantillon de données.

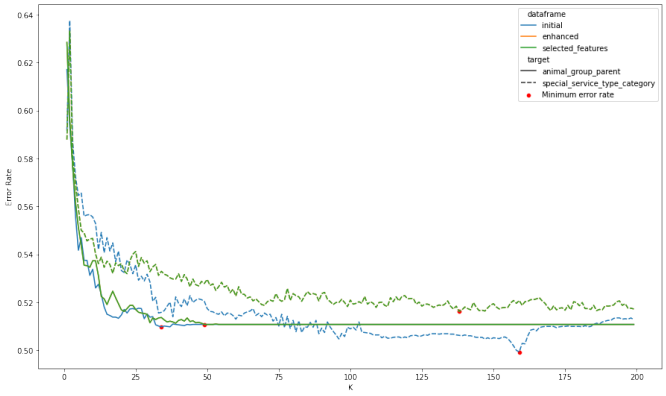

FIGURE 7 – Visualisation des taux d'erreur pour les échantillons en fonction de K

3.4 Comparaison des modèles et versions de jeu de données

Une fois les hyperparamètres trouvés, nous pouvons utiliser nos jeux de données d'apprentissage afin de comparer les différents modèles avec de la validation croisée. Nous prenons soin à cette étape de fixer la graine aléatoire de la cross-validation pour qu'elle soit similaire d'une méthode à l'autre.

Plusieurs observations sont à faire en voyant la comparaison des différents modèles et versions de jeu de données.

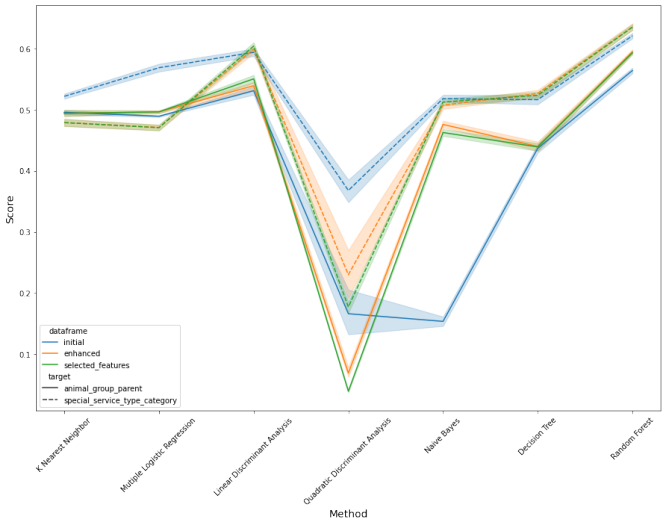

FIGURE 8 – Comparaison des scores des modèles

Premièrement, nous remarquons que les jeux de données *amélioré* et *variables sélectionnées* ont tendances à donner de meilleurs résultats que le jeu de donnée *initial*. C'est un bon signe, l'expertise géographique et temporelle apportée permet aux modèles de mieux classifier. Cela signifie aussi par ailleurs que notre sélection de variables n'a pas détériorée le jeu de données *amélioré*. Au contraire elle permet même d'améliorer le score. Dans un contexte de réchauffement climatique, il est judicieux de remarquer que nous avons réduit par 2 le nombre de variables tout en gardant la qualité de classification.

Ensuite nous pouvons remarquer c'est généralement la target *type d'intervention* qui obtient un meilleur score comparé à la target *type d'animal*. Cela peut déjà s'expliquer par le nombre de modalités différent (respectivement 4 et 10 modalités).

Enfin, nous pouvons noter que les méthodes Random Forest puis Linear Discriminant Analysis permettent d'obtenir les meilleurs scores pour les 2 targets. L'arbre de décision a déjà un score correct, la forêt d'arbres qui contient plusieurs arbres de décisions booste ces performances. La performance du modèle d'analyse discriminant linéaire peut se justifier par la distribution des classes qui est normale ainsi que par le fait que les classes sont similaires de par leur orientation et volume. L'analyse quadratique a un score beaucoup plus faible car elle ne fait que l'hypothèse de la distribution normale et que par conséquent le nombre de paramètres à trouver est beaucoup plus important.

3.5 Conclusion

Pour conclure, nous avons pu voir quelques analyses du jeu de données : visualisations spatiales, temporelles, ACP, CAH, K-moyennes... Nous avons pu distinguer quelques classes avec ces méthodes mais nous avons vu qu'elles ne correspondaient pas à nos attentes de classification.

La classification supervisée a quant-a-elle permis de créer et faire apprendre un classifieur sur la target désirée. Nous avons pu comparer différents modèles, jeux de données avec apport d'une expertise et sélection de variables. Nous avons alors obtenu de bons résultats que nous avons essayés de justifier en fonction des classifieurs.

Nous vous invitons également à voir notre code sur le repository Github SY09-Projet afin de voir l'ensemble des opérations effectuées sur ce jeu de données.

3.6 Pour aller plus loin

Nous avons décidé de ne pas plus pousser certains points qui ont été évoqués dans ce rapport pour plusieurs raisons : ressources humaines à allouer, manque d'expertise dans le domaine, hors cadre de l'UV et du projet de SY09.

Nous pouvons par exemple évoquer l'expertise apportée qui est intéressante mais qui aurait pu encore plus être exploitée. Nous aurions pu par exemple indiquer le nombre d'éléments présents à moins de x mètres, indiquer la superficie des éléments polygones, ajouter la longueur des éléments lignes... Par ailleurs pour plus de rigueur, il aurait fallu prendre en compte les éléments présents aussi autour de Londres (et non juste dans Londres), faire attention aux dates d'existence de ces éléments par rapport aux dates des interventions. Nous aurions également pu apporter d'autres informations comme la météo le jour de l'intervention, la luminosité (jour / nuit)...

Il aurait aussi été judicieux de voir si nous pouvions effectuer des transformations mathématiques sur certaines variables comme l'application de la fonction exponentielle, carré...

Les targets utilisées ici n'ont pas de réel sens pratique si nous voulions créer un outil d'aide aux brigades de pompiers de Londres. Nous aurions pu par exemple retirer des variables déterminées après une intervention (coût de l'intervention par exemple) puis faire un outil permettant de prédire la localisation de l'intervention puis le type d'animal et type d'intervention.

Il existe en effet des études faites sur la prédiction de lieu d'intervention (sur les crimes en l'occurrence) que nous pourrions exploiter. Une des implémentions se calque sur les modèles sismiques. Nous vous invitons à voir une vidéo de vulgarisation sur le sujet : Peut-on prédire les futurs crimes ? Fouloscopie.

Enfin, pour compléter le point qui vient d'être évoqué, nous aurions aussi pu nous intéresser à des modèles de classification prenant en compte les précédentes opérations dans la prédiction de nouvelles.

4 Annexes

4.1 Figures

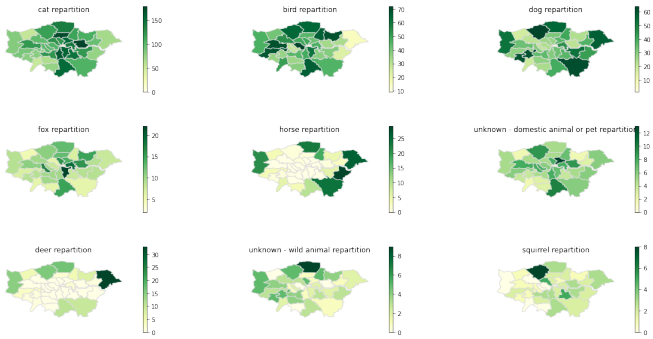

FIGURE 9 – Répartition des interventions sur les quartiers par type d'animal

Références

[1] *The London Fire Brigade rescues hundreds of creatures every year – from pigs to cats to bearded dragons.*, publié sur le site de la brigade des sapeurs-pompiers de Londres

[2] Perkin Amalaraj *Animal rescue costs in London up more than 50% since 2016*, publié le 16 février 2022 pour SW Londoner

[3] Constance Kampfner, *Animal rescues by London fire brigade rise 20% in pandemic year*, publié le 8 janvier 2021 pour The Guardian

[4] *London's firefighters attended two animal rescues a day in 2020*, publié le 7 janvier 2021 pour le site de la brigade des sapeurs-pompiers de Londres

[5] M. B. Short, M. R. D'Orsogna, V. B. Pasour, G. E. Tita, P. J. Brantiham, A. L. Bertozzi et L. B. Chayes, *A statistical model of criminal behavior*, publié le 28 décembre 2007