Compte rendu projet tutoré

Modélisation prédictive de la population canine en France à partir des données socio-démographiques des communes.

Réalisé par

Abdelkarim HADDAD
Cheick NIANG
Adam GBAGUIDI

UNIVERSITÉ PARIS 8 DES CRÉATIONS

Encadré par M. Vincent GODARD
Remerciement à M. Éric DELMELLE

Introduction

Notre étude a porté sur la prédiction du nombre de chiens par commune, en s'appuyant sur des variables telles que la population humaine et en utilisant une modélisation statistique appropriée. L'objectif de ce projet est d'établir la faisabilité de prédire le nombre de chiens dans une commune française sélectionnée aléatoirement, en appliquant un modèle statistique et en exploitant des variables pertinentes. Il s'agit donc de déterminer le modèle et les variables les plus adéquats pour répondre à notre problématique, formulée comme suit :

« Comment prédire le nombre de chiens par commune en fonction des caractéristiques démographiques de celles-ci? »

La première partie de ce rapport est dédiée à la présentation des données utilisées dans le cadre de ce projet, de leurs sources, ainsi que des différentes variables examinées. La deuxième partie se concentre sur le nettoyage, le traitement des valeurs manquantes et la production de variables utiles à l'entraînement du modèle, incluant la gestion des valeurs manquantes, l'ajout de nouvelles variables et la justification des choix méthodologiques et des modèles appliqués. La troisième partie expose le modèle final retenu pour la prédiction du nombre de chiens par commune, ainsi que les résultats obtenus. Enfin, nous conclurons ce compte rendu par une discussion relative aux limites de l'étude et par une ouverture sur d'éventuelles améliorations et perspectives futures.

Bien que ce travail s'inscrive dans le cadre d'un projet universitaire, son utilité peut être envisagée dans des contextes opérationnels, par exemple pour des entités privées telles que les chaînes de supermarchés. Lors de l'implantation d'un nouveau point de vente dans une commune, l'estimation du nombre de chiens présents peut permettre d'anticiper la demande en produits pour animaux, notamment la nourriture pour chiens. Ceci peut contribuer à la détermination de l'espace à allouer à ces produits en magasin, dans le but d'optimiser l'agencement des rayons et d'éviter des situations de surstock ou de rupture. Ainsi, la prédiction du nombre de chiens peut constituer un outil d'aide à la décision pour une meilleure gestion de l'espace et des ressources.

Partie 1 : Lecture des données

1.1 Source des données

Pour réaliser ce travail, nous disposons d'un jeu de données de l'ICAD (Identification des carnivores domestiques), organisme officiel chargé de l'identification des chats, chiens et

furets domestiques en France. Ce jeu de données nous est particulièrement utile, car il recense le nombre de chiens présents dans chaque commune française, et ce, pour différentes années.

Dans ces données, nous avons donc le nombre de chiens par année, de 2013 à 2020, dans chaque commune de France (36 594 communes). Ce nombre de communes ne correspond pas exactement au nombre de communes d'aucune des années entre 2013 et 2020. Il s'agit du résultat d'une agrégation de plusieurs tableaux de longueur differentes, conséquence naturelle des réaménagements administratifs en France. Ce qui entraîne quelques décalages et de fausses valeurs manquantes.

Nous disposons également du nombre d'habitants pour chaque commune et pour chaque année sur la même période (2013-2020). Le nombre d'habitants est une variable primordiale pour ce projet, car notre hypothèse est que nous pouvons prédire le nombre de chiens en fonction du nombre d'humains dans une commune.

Afin de confirmer cette hypothèse, nous avons réalisé un scatter plot représentant la corrélation entre la population humaine et la population canine par commune. Ce graphique met clairement en évidence une forte corrélation positive entre ces deux variables, confirmant que les communes les plus peuplées tendent également à compter un plus grand nombre de chiens.

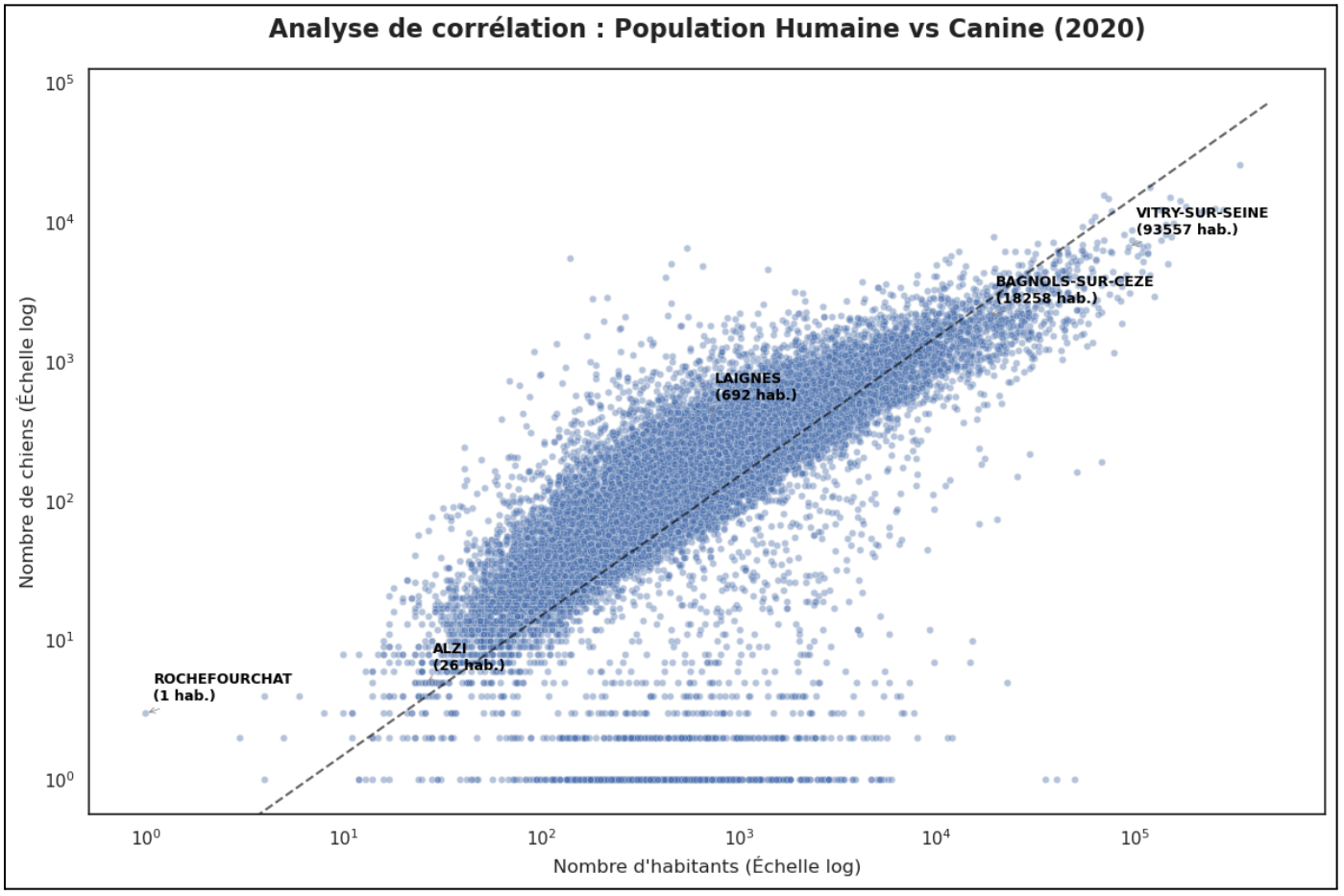

Figure 1 : nuage de point correlation variables population chien population humaine échelle log

1.2 Description des colonnes :

Dans un premier temps, nous disposons de plusieurs colonnes donnant des informations géographiques et administratives sur chaque commune. Leur détail est donné dans l'annexe 1. Celles-ci comprennent la superficie, les coordonnées du centroïde et du chef-lieu de la commune ainsi que l'altitude moyenne, en plus du département et de la région d'appartenance de la commune.

Les colonnes qui concernent le comptage de la population de chiens sont notées Po_chien13 à Po_chien20 : nombre de chiens recensés dans la commune pour chaque année, de 2013 à 2020. Il en va de même pour la population humaine avec les colonnes Pop_huma13 à Pop_huma20.

1.3 Données manquantes

Dans le jeu de données, certaines colonnes de comptage présentent des valeurs manquantes ou des valeurs égales à zéro, notamment pour 'Po_chien', où l'on retrouve en moyenne environ 30 % des colonnes dans un état de NaN (Not a Number). Ces valeurs peuvent correspondre à des failles dans la remontée d'information par les vétérinaires, ou à des erreurs dans le traitement et l'agrégation des données ou meme,mais cela est moins probable à une véritable absence d'enregistrement de chiens, en effet les zéros dans ce jeu de données sont assimilé à des NaN. On observe également la présence de valeurs Po_chien qui apparaissent comme non cohérentes au regard des ordres de grandeur observés pour les autres communes et des taux de chiens pour humains à l'échelle nationale (0,15 chien par humain), notamment environ 800 communes avec un ratio chiens/humains anormalement bas. D'autres, au contraire, présentent des taux de chiens par humain très élevés (15 chiens/humain). Dans la suite de l'étude, nous examinerons ces incohérences et évaluerons si ces dernières intègrent les données d'entraînement du modèle ou si ces valeurs seraient assimilées à des données manquantes. Les différentes hypothèses liées à ces valeurs, ainsi que les choix de traitement associés, seront détaillées dans la partie consacrée au nettoyage et au formatage des données.

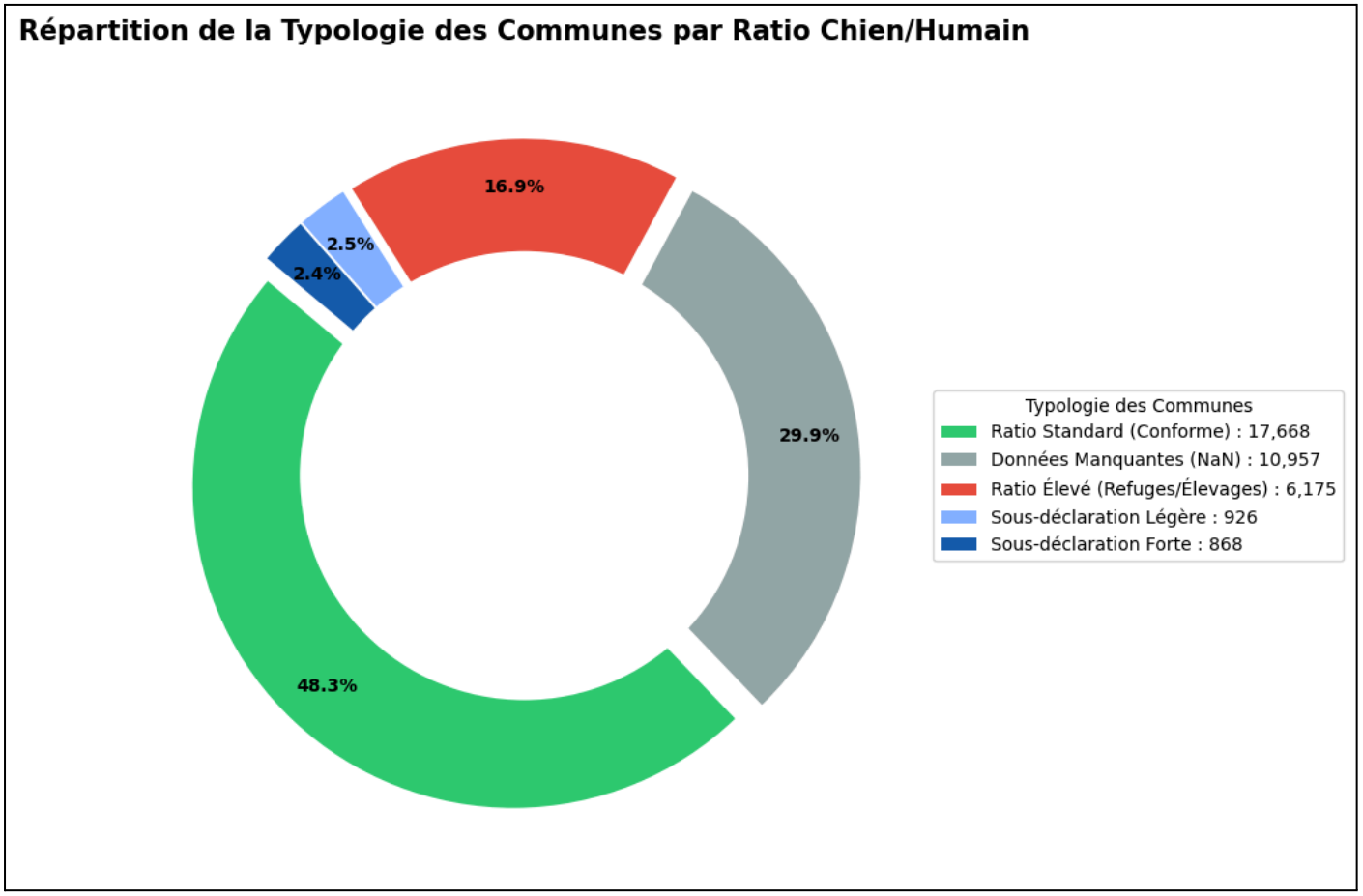

Figure 2 - Typologie des anomalie des données

Pour enrichir nos données et rendre notre modèle plus fiable, nous avons récupéré des données supplémentaires de l'INSEE contenant des informations sur la structure des ménages dans les communes et la typologie des communes.

Avant de procéder au nettoyage et à la modélisation, nous avons analysé la répartition des principales variables à l'aide d'histogrammes, notamment la population humaine et la population canine par commune pour l'année 2020. La distribution de la population humaine montre une forte asymétrie : la majorité des communes françaises comptent peu d'habitants, tandis qu'un nombre restreint de communes concentre des populations beaucoup plus élevées. Cette structure reflète la forte proportion de communes rurales en France. La distribution du nombre de chiens par commune présente une forme similaire, avec une majorité de communes ayant un nombre relativement faible de chiens, et quelques communes affichant des valeurs nettement plus élevées. Ces histogrammes mettent en évidence une forte hétérogénéité des données et confirment l'existence de fortes disparités entre communes, ce qui justifie la nécessité d'un travail approfondi de nettoyage et l'utilisation de modèles capables de s'adapter à cette variabilité.

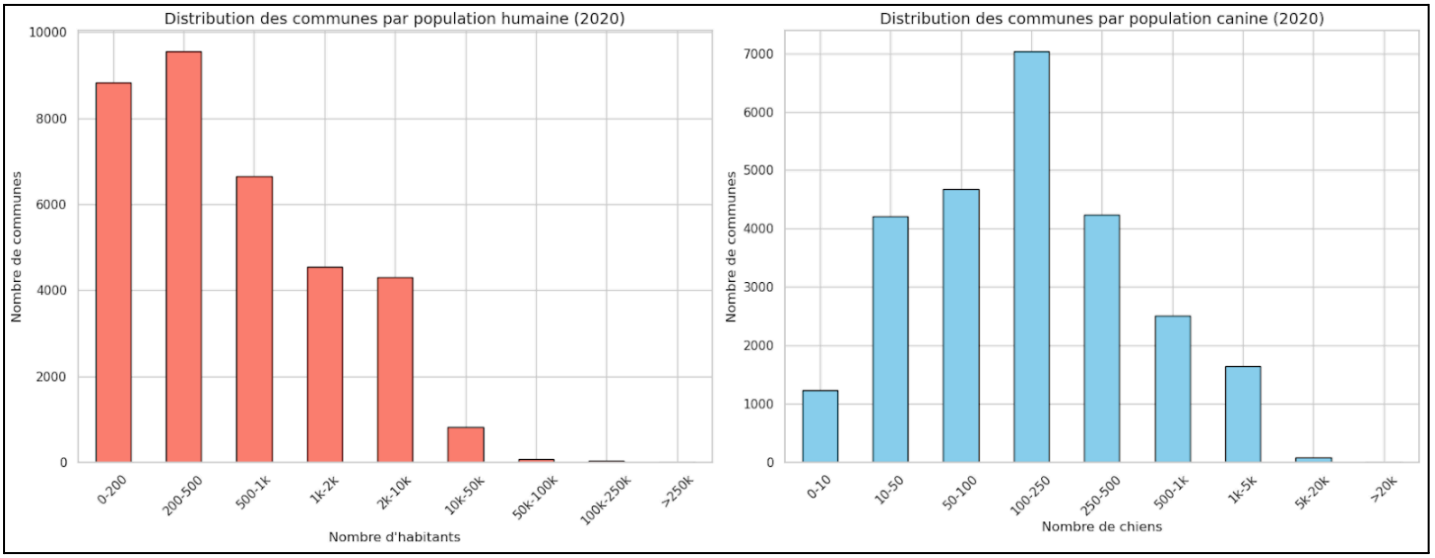

Figure 3 - Histogramme distribution de la population humaine et la population de chien en 2020

1.4 Disposition et typologie des valeurs manquantes et des valeurs aberrantes

Au-delà du volume important de données manquantes (près de 11 000 communes en NaN), la compréhension de la distribution de la population canine nécessitait la création d'une variable normalisée : le ratio Chiens/Humain (ratio_ch_hum). Cette métrique permet de neutraliser l'effet de taille des communes (superficie et population) pour offrir un indicateur de densité relative.

L'exploration de cette variable a mis en lumière une forte hétérogénéité, avec des valeurs oscillantes entre l'invraisemblable (quasi-nulles) et l'aberrant (maxima à 38 chiens/habitant), alors que la moyenne nationale issue de l'I-CAD se situe autour de 0,15 (soit environ 1 chien pour 8 habitants). Pour qualifier ces anomalies, nous avons appliqué une méthodologie de détection basée sur un écart interquartile (IQR) hybride centré sur la moyenne nationale, couplée à une analyse spatiale (LISA) pour identifier les clusters géographiques.

L'analyse comparative des attributs sociodémographiques nous permet de dresser le "portrait-robot" de ces trois catégories de communes atypiques :

1. Les Communes « Sans Données » (NaN) : La marque de l'hyper-ruralité

Les communes absentes de la base (NaN) ne sont pas reparties aléatoirement. Elles correspondent au profil type de la micro-commune rurale.

- **Démographie** : Elles sont extrêmement peu peuplées (médiane de 233 habitants contre 669 pour la norme, soit -65 %).
- **Urbanisation** : Elles présentent une très faible artificialisation des sols (1,4 % contre 4,7 %) et une densité de population moitié moindre que la norme (-55 %).
- **Géographie** : Ce manque de données semble structurel, lié à des zones de très faible densité où le maillage vétérinaire peut être moins dense. Géographiquement, ces communes se concentrent particulièrement en Normandie et dans les Pays de la Loire.

2. Les Communes en « Sous-déclaration » : Une anomalie administrative

Contrairement aux NaN, les communes présentant un ratio quasi-nul ne se distinguent pas structurellement des communes standards.

- Profil Standard : Leur population (649 hab.), leur densité, la taille des ménages et le revenu médian sont quasiment identiques à la moyenne nationale (écarts inférieurs à 3 %).
- Interprétation : L'absence de chiens enregistrés dans ces communes ne s'explique pas par la géographie ou la sociologie, mais relève probablement d'un défaut de remontée d'information ou d'une anomalie administrative ponctuelle, car rien ne les prédispose à une absence totale de canidés.

3. Les Communes en « Surtaxe » (Ratios élevés) : Terres d'élevage et de chasse

Les communes affichant des ratios (chien/humain) très élevés (+127 % par rapport à la norme) dessinent le portrait d'une ruralité profonde et vaste.

- **Espace et Faible Densité** : Ce sont des communes vastes (+23 % de superficie) mais très peu peuplées (303 habitants en médiane), résultant en une densité de population extrêmement faible (21 hab/km², soit -62 % par rapport à la norme).
- **Sociologie** : Avec un revenu légèrement inférieur à la médiane (-5,4 %) et une très faible artificialisation, ces zones correspondent à des territoires agricoles ou forestiers propices à l'accueil de structures nécessitant de l'espace (élevages, refuges, meutes de chasse).
- **Géographie** : L'analyse spatiale confirme cette hypothèse en révélant des clusters marqués dans des régions à forte tradition rurale : Centre-Val de Loire, Bretagne et Nouvelle-Aquitaine.

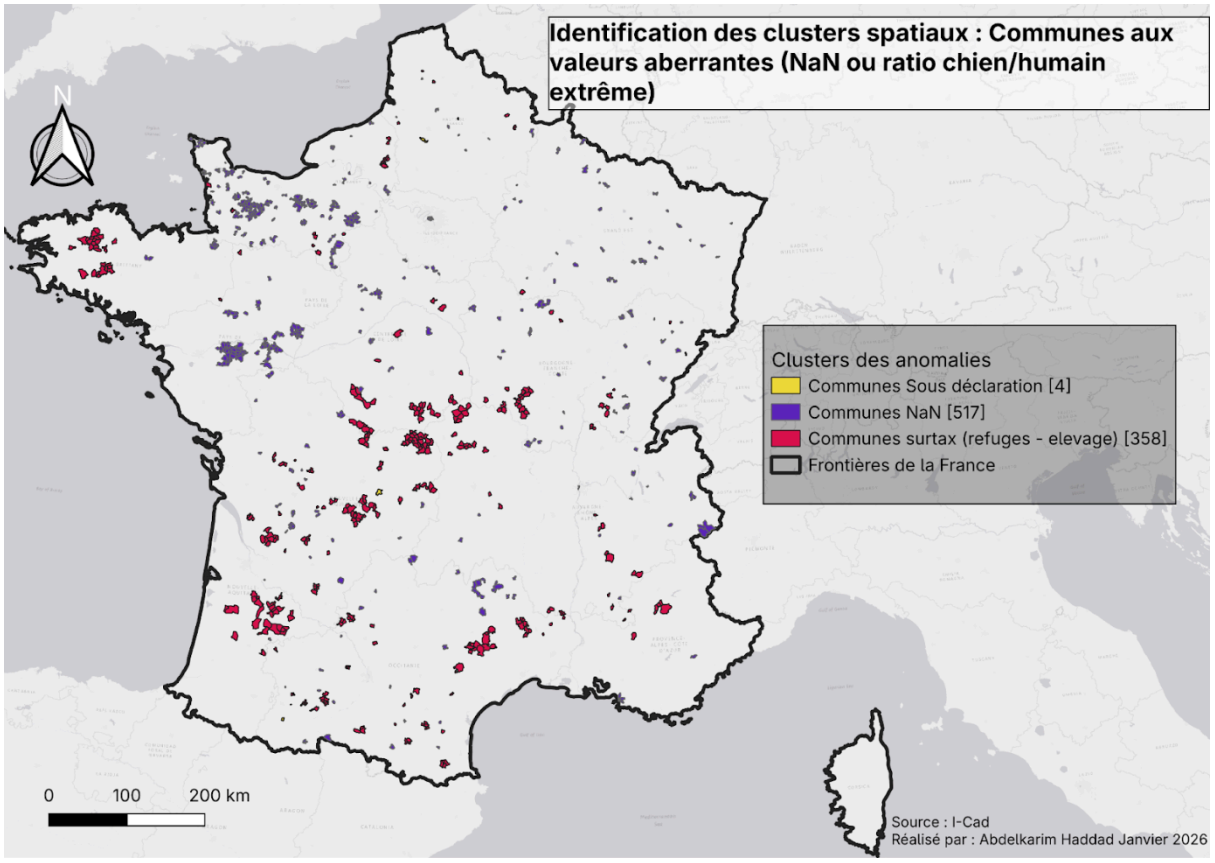

Fig. 4 - Identification des cluster spatiaux des valeurs manquantes et aberrantes

Partie 2 : Traitements des données

2.1 Gestion des valeurs manquante et imputation temporelle

Quand on s'intéresse au NaN sur les 8 années, on constate que pour toute les année (environ 11 000 communes, soit 30 %) des communes sont renseigné comme NaN cela dit on distingue deux types de commune Celle où il y a aucun enregistrement sur les 8 ans, et d'autre ou on a quand même quelque enregistrement ponctuels.

Po_chien13	Po_chien14	Po_chien15	Po_chien16	Po_chien17	Po_chien18	Po_chien19
None Count 2621	None Count 2518	None Count 2587	None Count 2624	None Count 2606	None Count 2692	None Count 2715
NaN	NaN	135.0	101.0	NaN	NaN	115.0
238.0	251.0	263.0	261.0	267.0	260.0	NaN
NaN	NaN	497.0	493.0	454.0	510.0	598.0
NaN	NaN	NaN	NaN	440.0	NaN	NaN
NaN	NaN	NaN	94.0	NaN	NaN	73.0
NaN	NaN	NaN	2.0	155.0	155.0	NaN
NaN	1.0	1.0	1.0	221.0	1.0	253.0
57.0	51.0	49.0	57.0	62.0	63.0	56.0
59.0	63.0	1.0	NaN	57.0	55.0	55.0
107.0	114.0	120.0	NaN	NaN	NaN	NaN
NaN	NaN	36.0	NaN	NaN	NaN	NaN
4.0	6.0	8.0	8.0	12.0	15.0	15.0
NaN	NaN	39.0	NaN	40.0	39.0	NaN
NaN	NaN	NaN	NaN	222.0	NaN	NaN
NaN	1.0	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	26.0	NaN	21.0	NaN
34.0	NaN	NaN	40.0	NaN	NaN	NaN
139.0	171.0	331.0	323.0	377.0	375.0	432.0
NaN	81.0	NaN	NaN	1.0	117.0	1.0
NaN	NaN	8.0	6.0	12.0	18.0	18.0
NaN	NaN	NaN	NaN	NaN	5.0	NaN
NaN	NaN	NaN	NaN	NaN	265.0	257.0
79.0	87.0	NaN	NaN	NaN	NaN	NaN
NaN	43.0	NaN	45.0	48.0	52.0	NaN
NaN	NaN	NaN	62.0	55.0	72.0	65.0
NaN	NaN	1.0	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN	NaN	111.0

Figure 5 - Répartition et persistance des communes sans enregistrement (NaN) sur 8 ans

Comme le démontre le tableau ci-dessus, de nombreuses communes présentent des enregistrements manquants ou partiels sur certaines années, elles sont d'un nombre suffisamment important (10957 communes) pour fragiliser les résultats de l'entraînement des modèles. La performance d'un modèle comme le Random Forest dépend directement du volume et de la qualité des données d'entrée.

Pour pallier ce manque sans introduire de biais statistique, nous avons mis en place un algorithme d'imputation temporelle. Son principe repose sur une hypothèse de stabilité sociologique : le comportement de possession canine (le ratio chiens/habitants) d'une commune évolue peu à court terme, même si sa population humaine fluctue.

Même si la majorité des NaN (6594/10957) sont constant dans le temps il reste une proportion non négligeable (4363) des NaN qui peuvent bénéficier d'un traitement d'imputation temporelle

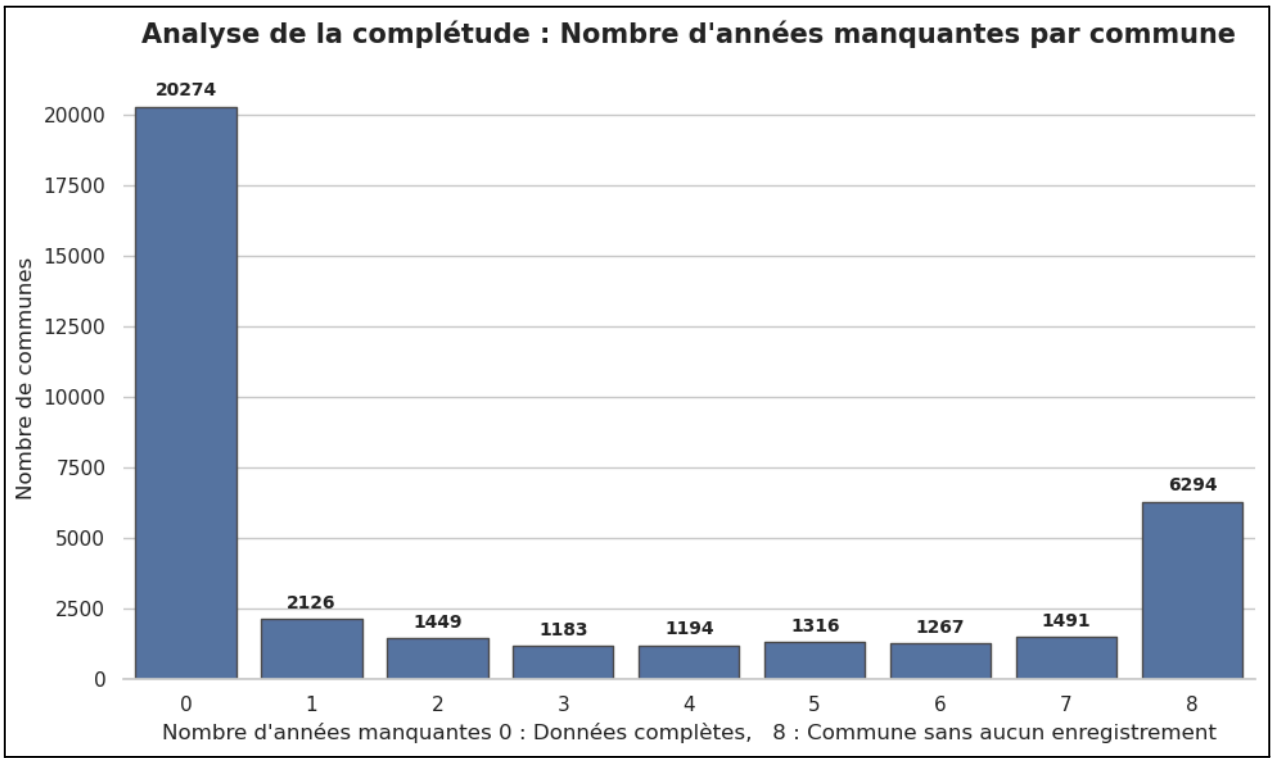

Figure 6 - Distribution du nombre d'années manquantes par commune (2013–2020)

2.1.1 Méthodologie : L'imputation temporelle

Le processus se décline en trois phases de calcul et de filtrage :

A. Phase de Cadrage (Pré-calculs)

Pour chaque commune et chaque année où la donnée est disponible, nous calculons le ratio de possession de chien (TxCh_hum)

Afin de distinguer les données fiables des anomalies, nous définissons des seuils d'acceptabilité basés sur :

1. La moyenne départementale des ratios.
2. Un intervalle de fiabilité basé sur l'IQR (écart interquartile).

B. Phase de Filtrage et Sélection

L'algorithme analyse ensuite chaque commune pour déterminer son éligibilité à l'imputation

- Cas des "All-NaN" (6 294 communes) : Si aucun enregistrement n'existe sur toute la période (2013-2020), la commune est exclue de l'imputation temporelle.
- Cas des Communes Partielles (4 663 communes) : Si la commune possède au moins un enregistrement historique, l'algorithme vérifie si ce ratio “TxCh_hum” est compris dans l'intervalle acceptable défini précédemment.

C. Phase d'Imputation ("Fallback")

Lorsqu'une donnée est manquante pour l'année cible (2020), l'algorithme parcourt l'historique de la commune :

1. Il sélectionne la valeur valide la plus récente (ex: 2020, sinon 2019, sinon 2018...).
2. Il Prend la valeur de population canine de l’année retenue et applique une correction qui tient compte de l’évolution de la population humaine qui sépare 2020 et l’année en question

   Exemple :

$$Population\ Chien\ Final\ = Population\ Chien\ 2019 \times \frac{Population\ Humaine\ 2020}{Population\ Humaine\ 2019}$$

Sur un total de 36 594 communes analysées, la méthodologie a conduit à la correction/imputation de **2 644** communes, tandis que 26 409 autres ont été validées sans intervention. Cela signifie que 7 541 communes (soit 20,6 %) sont considérées comme invalides selon les critères définis.

2.1.2 Spatial Lag

Au-delà des variables sociodémographiques classiques, nous avons intégré le décalage spatial (spatial lag) du ratio chiens/habitants pour saisir le contexte local et les effets de voisinage.

Le Spatial Lag est une moyenne pondérée des valeurs d'une variable chez les voisins. Conformément à la première loi de Tobler (« les choses proches sont plus liées »), il permet au modèle de capturer l'effet de similarité régionale (ex: une commune aura probablement beaucoup de chiens si ses voisins en ont).

Le choix du ratio (normalisé) plutôt que du nombre brut de chiens (Po_chien) est stratégique, il isole la "propension culturelle à posséder un chien" de la démographie. Un lag sur le nombre brut aurait introduit un biais de taille important et risquait une fuite de données (data leakage), le modèle apprenant indirectement la cible par la population environnante.

2.1.3 Sélection du modèle statistique approprié

Dans un premier temps, nous avons testé le modèles de référence pour les données de comptage classiques comme le modèle de Poisson, car notre variable à prédire correspond à un nombre de chiens par commune. Cependant, ce modèle s'est vite révélé peu pertinent dans notre cas, car nos données présentent une variance beaucoup trop importante (450295.8) : certaines communes ont très peu de chiens, tandis que d'autres en comptent plusieurs milliers, ce qui crée une dispersion très forte. Le modèle de Poisson, qui suppose une variabilité plus "stable", a donc tendance à mal s'adapter à ce type de données et à produire des prédictions moins fiables.

Nous avons alors essayé un modèle binomial négatif, un autre GLM, qui est censé mieux gérer ce problème de dispersion, mais les résultats restaient limités. Dans notre cas, la relation entre les variables explicatives et le nombre de chiens semble plus complexe et pas forcément linéaire, ce qui fait que ce modèle ne parvenait pas à exploiter correctement toutes les variables disponibles, ni à capter certaines différences importantes entre les communes. C'est donc pour ces raisons que nous nous sommes tournés vers un modèle Random Forest, plus flexible, capable de mieux prendre en compte des relations non linéaires et l'hétérogénéité observée entre les communes.

Le Random Forest est un modèle basé sur un ensemble d'arbres de décision. Au lieu de construire un seul arbre, on en construit plusieurs, chacun entraîné sur un échantillon un peu différent des données et avec un choix aléatoire de variables. Ensuite, on regroupe toutes les prédictions des arbres pour obtenir une prédiction finale plus stable, ce qui est utile quand les données sont très dispersées et que les relations entre variables ne sont pas forcément linéaires.

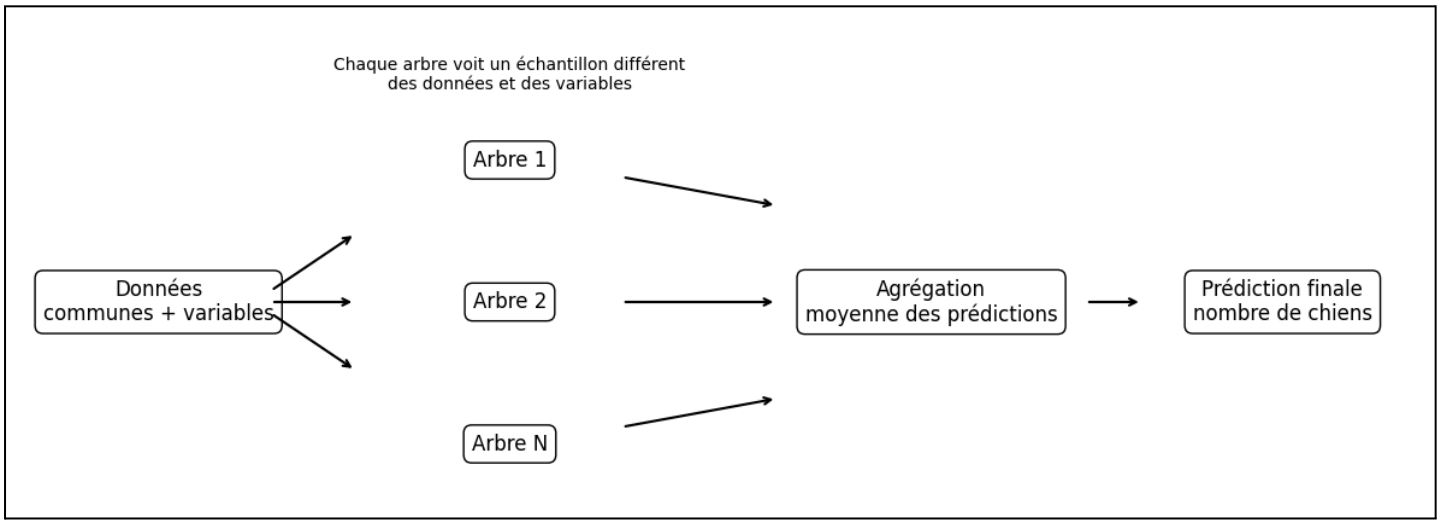

Figure 7 - Schéma fonctionnement Random Forest

Partie 3 : Analyse des Résultats

3.1 Performance du modèle prédictif Random Forest

Pour modéliser la population canine, nous avons entraîné un modèle de Random Forest sur un jeu de données enrichi, incluant des variables sociodémographiques (revenus, types de logements, composition familiale) et des variables plus contextuelles, comme le spatial lag du ratio chiens humains.

Globalement, le modèle obtient de très bonnes performances, avec un R^2 de 0,90 et une erreur médiane MAE d'environ 38 chiens par commune sur le jeu de test. L'analyse de l'importance des variables met en évidence une hiérarchie assez nette. La population humaine Pop_huma20 est de loin la variable la plus importante, ce qui confirme le lien évident entre démographie humaine et présence canine. Le contexte local, via ratio_lag_final, ainsi que la superficie, jouent ensuite un rôle important pour corriger les prédictions, notamment dans les communes plus rurales ou atypiques.

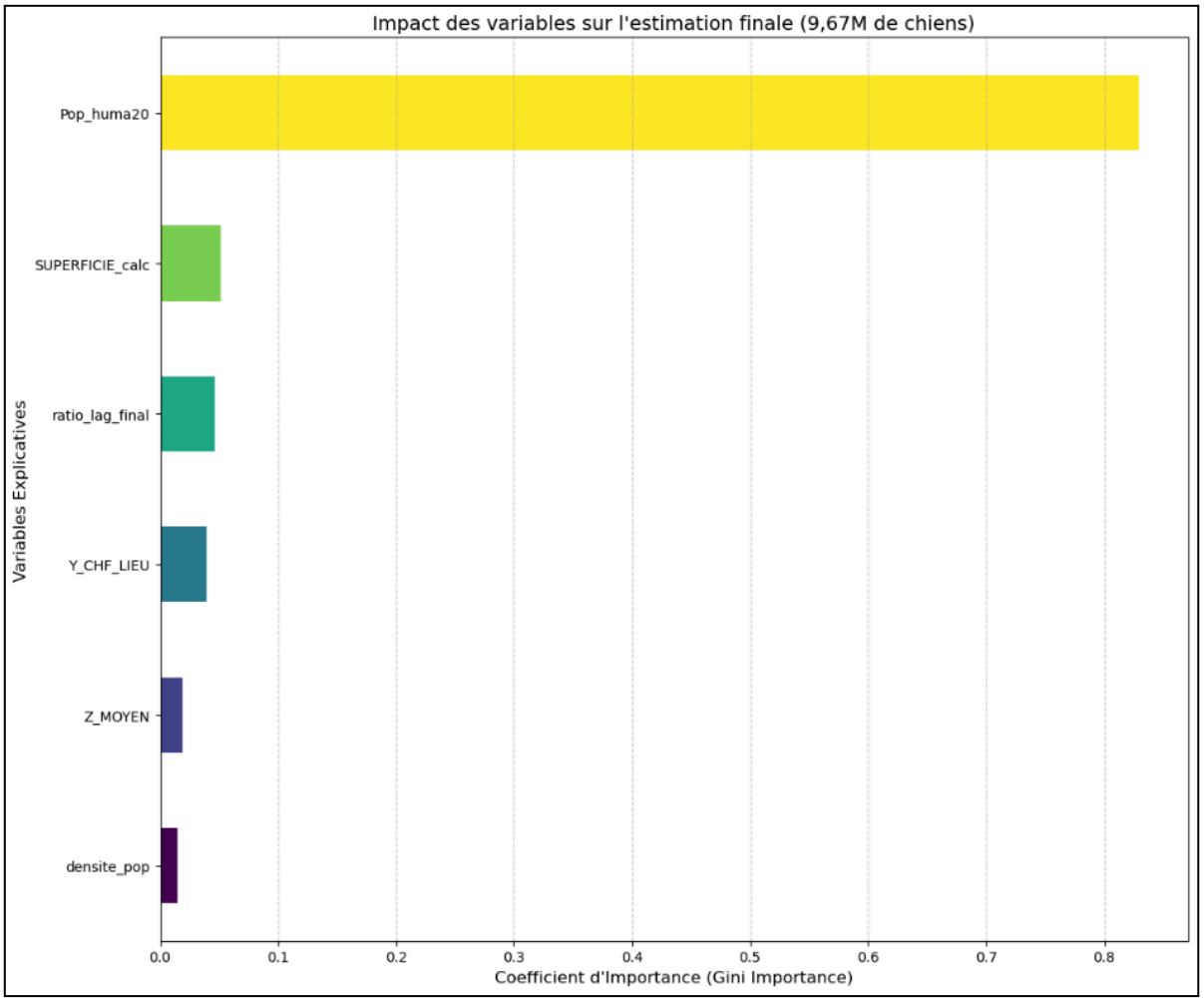

Figure 8 - Contribution des variables à la prédiction dans le modèle Random Forest.

Ces résultats doivent toutefois être nuancés. Le score R^2 est en partie tiré vers le haut par les communes très peuplées, qui pèsent fortement dans la variance totale. Une analyse par strates montre que le modèle est particulièrement robuste pour les communes moyennes et grandes, mais que la variance reste plus difficile à expliquer pour les micro-communes, notamment celles de moins de 2 000 habitants.

Un point assez surprenant est que les variables “sociologiques” et d’occupation ajoutées pour décrire la typologie des communes ont finalement eu très peu d’impact sur la performance du modèle. Des variables comme la part de maisons, le nombre de familles, la part de familles avec deux enfants ou plus, la part des surfaces artificialisées, le revenu médian, la taille moyenne des ménages, la typologie des ruralités ou encore la présence d’un refuge, n’apportent qu’un gain marginal.

Au final, seules quelques variables dépassent réellement 1 % d’importance dans le modèle, notamment Population humain de 2020, densité de population, latitude chef lieu, Altitude, spatial lag du ratio et superficie. Cela montre que les facteurs structurels et géographiques dominent largement, et que la performance du modèle repose surtout sur un ensemble démographie, morphologie et contexte local. Les variables socio-économiques plus fines, même si elles semblent pertinentes sur le papier, apportent peu d’information supplémentaire, probablement parce qu’une partie de leur effet est déjà captée indirectement par la densité et par le voisinage spatial. Cela étant dit, on aurait pu s’attendre à ce que certaines variables liées au type d’habitat, comme la typologie des ruralités ou le niveau d’artificialisation, aient un effet plus marqué.

3.2 Analyse des Résidus : Limites et Robustesse

L'analyse fine des erreurs du modèle (résidus) révèle deux phénomènes opposés :

- L'effet "Plafond de Verre" métropolitain : Le modèle a tendance à surestimer le nombre de chiens dans les très grandes métropoles (Paris, Lyon, Marseille). Il peine à modéliser l'effet dissuasif de l'hyper-densité urbaine sur la possession animale.
- Le paradoxe des micro-villages : Sur certaines petites communes, le modèle prédit des valeurs sociologiquement "normales" (ex: 50 chiens pour 500hab) alors que la donnée I-CAD qu’on a enregistré des valeurs extrêmement faibles (ex: 3 chiens). Cela engendre des erreurs relatives (%) artificiellement élevées. Dans ces cas précis, nous considérons que la prédiction du modèle est probablement plus proche de la réalité sociologique que la donnée administrative brute, potentiellement lacunaire.

Malgré ces écarts, la médiane des résidus se situe à -1,12, indiquant que pour la majorité des communes françaises, le modèle prédit le nombre de chiens à une unité près.

3.3 Validation Spatiale (Test de Moran)

Afin de vérifier si le modèle avait omis des logiques régionales (biais géographique), nousavons effectué un test d'autocorrélation spatiale de Moran sur les résidus.

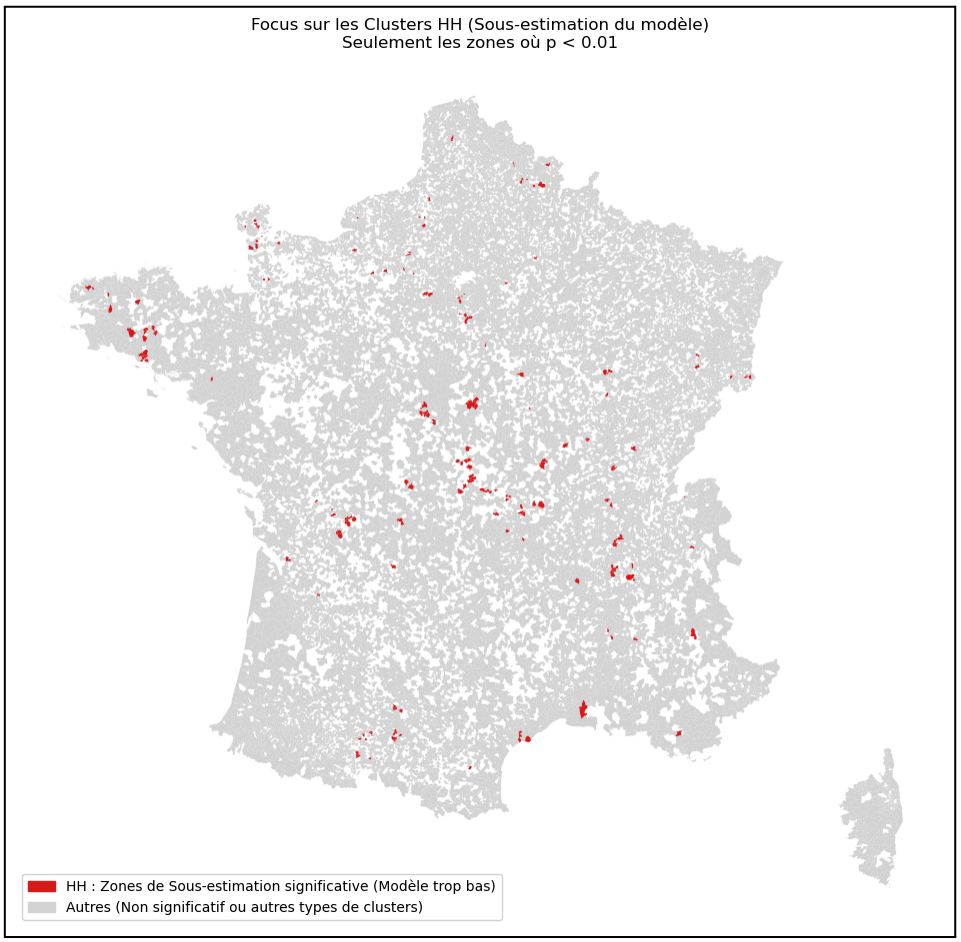

Figure 8. Clusters locaux des résiduels selon la méthode LISA

- Indice de Moran Global : **0,0090**
- P-value : **0,002**

Ce résultat, extrêmement proche de zéro, indique une distribution aléatoire des erreurs. Contrairement à une approche linéaire classique qui aurait pu biaiser des régions entières, notre modèle ne présente aucune structure spatiale résiduelle. L'intégration de la variable de Spatial Lag a permis de capturer efficacement les spécificités locales. Les erreurs restantes sont du "bruit blanc", validant ainsi la robustesse géographique du modèle à l'échelle nationale

Conclusion

L'objectif de ce projet était de déterminer dans quelle mesure il est possible de prédire le nombre de chiens par commune à partir de variables disponibles, en particulier la population humaine. À partir des données I-CAD sur la période 2013–2020, nous avons commencé par explorer la structure des données et confirmé une relation forte entre population humaine et population canine. Ce travail exploratoire a également mis en évidence un volume important de valeurs manquantes et des comportements non aléatoires, notamment des communes sans enregistrement sur toute la période, ce qui suggère un biais structurel plutôt qu'un simple bruit de saisie.

Pour limiter l'impact de ces manques, nous avons mis en place une stratégie de traitement combinant une imputation temporelle lorsque l'historique de la commune le permettait, et l'ajout d'une variable de contexte local via le spatial lag du ratio chiens habitants. Les modèles de comptage classiques, comme Poisson ou binomiale négative, se sont révélés peu adaptés à cause de la forte dispersion et de la complexité des relations entre variables. Nous avons donc retenu un modèle Random Forest, plus flexible, qui obtient de bonnes performances globales, avec un R^2 de 0,90 et une MAE médiane d'environ 38 chiens sur le jeu de test. L'importance des variables montre que la prédiction repose surtout sur des facteurs structurels, d'abord la population humaine, puis des variables de morphologie et de contexte local.

Enfin, l'analyse des erreurs met en évidence des limites attendues : le modèle a plus de mal sur les micro-communes et tend à surestimer certaines très grandes villes, ce qui suggère que des effets urbains spécifiques ne sont pas entièrement capturés. Malgré cela, les tests spatiaux sur les résidus indiquent que le modèle ne laisse pas de biais géographique majeur, ce qui renforce la robustesse de l'approche à l'échelle nationale. En perspective, une amélioration naturelle serait d'utiliser des données I-CAD plus complètes et plus homogènes, et d'intégrer des variables décrivant mieux l'environnement urbain et l'accès aux espaces extérieurs afin de mieux modéliser les grandes métropoles et les communes très atypiques.