



Estimation de la densité d'espèces sur des cartes géospatiales

Par
Groupe 2
* * * * *

Avodagbe Ze Paul-Valéry
Marah Djafar El Noumery
Ngessi Fils Michel Julien
Njoupouo Gnigni Samira Alima
Takou Nyabeye Mégane Fridile

Rapport soumis pour la validation de l'unité d'enseignement :
Information et Estimation des densités

Sous la supervision de
Dr Malong Yannick

Chargé de cours
École Nationale Supérieure Polytechnique
Université de Douala

Table des matières

Liste des tableaux	2
Liste des figures	3
Introduction	2
1 Fondements Théoriques de l'Estimation de Densité	3
1.1 Définition et enjeux de l'estimation de densité	3
1.2 Présentation des principales méthodes d'estimation de densité	3
1.2.1 L'histogramme	3
1.2.2 L'estimation par noyau	4
1.2.3 L'estimation paramétrique	4
1.3 Comparaison entre les méthodes et critères de choix	5
2 Étude de cas : estimation de la densité du paresseux à gorge brune et le rat de riz forestier via le KDE	6
2.1 Description des données utilisées	6
2.2 Choix de la méthode d'estimation de densité : Justification du KDE	7
2.3 Détails de l'implémentation	7
2.4 Visualisation des résultats : Cartes de densité	8
3 Analyse et Interprétation des Résultats	11
3.1 Interprétation des cartes de densité	11
3.2 Discussion de l'impact du paramètre de lissage (bande passante)	11
3.3 Pertinence pour l'écologie et la conservation de la biodiversité	13
3.4 Limitations de la méthode et des données	13
Conclusion	14
Annexes	16

Liste des tableaux

1.1 Comparaison entre histogramme, KDE et estimation paramétrique	5
---	---

Liste des figures

2.1	le paresseux à gorge brune (<i>Bradypus variegatus</i>)	6
2.2	le rat de riz forestier (<i>Microryzomys minutus</i>)	6
2.3	Distribution spatiale des espèces	7
2.4	Estimation de la densité des espèces via le noyau gaussien	9
2.5	Estimation de la densité des espèces via le noyau Epanechnikov	9
2.6	Estimation de la densité des espèces via le noyau rectangulaire	10
3.1	sous-lissage avec un noyau Gaussien (bande passante =0.01)	12
3.2	sur-lissage avec un noyau Gaussien (bande passante =0.09)	12

Introduction

L'estimation de densité est une technique statistique fondamentale qui a pour objectif d'approcher la fonction de densité de probabilité d'une variable aléatoire à partir d'un ensemble de données observées. Contrairement aux méthodes paramétriques qui reposent sur des hypothèses concernant la distribution sous-jacente des données (telles que la distribution normale ou exponentielle), l'estimation de densité propose une approche plus flexible et adaptative pour la modélisation des distributions de données réelles. Cette technique revêt une importance capitale en analyse de données, car elle permet notamment de mieux comprendre la répartition des valeurs, d'explorer les propriétés d'un ensemble de données, et de révéler des caractéristiques telles que l'asymétrie et la multimodalité. L'estimation de densité est également essentielle pour la détection d'anomalies, l'identification de structures cachées, et peut améliorer les performances en apprentissage automatique en prétraitant les données.

Ce rapport se concentre sur l'application de l'estimation de densité au projet n°2, qui vise à appliquer l'estimation de densité par noyau (KDE) sur des données de présence d'espèces pour estimer leur distribution géographique. L'objectif principal est de modéliser et de cartographier la probabilité de présence des espèces en fonction de facteurs environnementaux. Ce projet est d'une grande pertinence pour l'écologie et la conservation de la biodiversité, car la compréhension de la distribution spatiale des espèces est cruciale pour les efforts de surveillance des écosystèmes et des espèces, ainsi que pour la prise de décisions en matière de gestion de la conservation.

Ce rapport a pour objectifs de :

- Présenter les méthodes utilisées pour estimer la densité d'espèces à partir de données de présence, en mettant l'accent sur le KDE.
- Analyser et interpréter les résultats obtenus sous forme de cartes de densité, illustrant la distribution géographique estimée des espèces.
- Discuter des implications écologiques et pour la conservation de ces estimations de densité, en soulignant leur utilité pour la surveillance des habitats et la gestion des espèces.

Afin d'atteindre ces objectifs, ce rapport présentera dans un premier temps les fondements théoriques de l'estimation de densité et de la méthode du noyau (KDE). Ensuite, il détaillera l'application de cette méthode à l'estimation de la densité d'espèces sur des cartes géospatiales, suivie de l'analyse et de l'interprétation des résultats cartographiques. Enfin, une discussion portera sur les implications écologiques et les perspectives pour la conservation découlant de cette étude.

Fondements Théoriques de l'Estimation de Densité

1.1 Définition et enjeux de l'estimation de densité

L'estimation de densité est une technique statistique qui permet d'approcher la fonction de densité de probabilité d'une variable aléatoire à partir d'un ensemble de données observées. Spécifier cette fonction fournit une description naturelle de la distribution de la variable aléatoire et permet de calculer des probabilités associées. L'estimation de densité se concentre sur la construction d'une estimation de cette fonction de densité à partir des données disponibles.

Les enjeux principaux de l'estimation de densité sont multiples :

- **Visualisation de la distribution sous-jacente des données** : Elle permet une investigation informelle des propriétés d'un ensemble de données et peut révéler des caractéristiques telles que l'asymétrie et la multimodalité. Les estimations de densité sont idéales pour la présentation des données à des non-mathématiciens en raison de leur compréhensibilité.
- **Détection d'anomalies ou de valeurs aberrantes** : Les régions de faible densité dans l'espace des données peuvent indiquer des observations inhabituelles.
- **Identification des modes et des structures cachées** : L'estimation de densité peut aider à identifier les modes (maxima locaux) dans la distribution, ce qui peut correspondre à des clusters dans les données.

1.2 Présentation des principales méthodes d'estimation de densité

1.2.1 L'histogramme

L'histogramme est une méthode d'estimation de densité qui divise la plage des données en intervalles (ou classes) et estime la densité dans chaque intervalle en fonction du nombre d'observations qui y tombent. Formellement, si l'on a une dissection de la droite réelle en bins, l'estimation est définie par le nombre d'observations tombant dans chaque bin, divisé par la largeur du bin et le nombre total d'observations.

Le choix des intervalles, notamment leur largeur, a une influence significative sur la représentation de la densité. Des intervalles trop larges peuvent masquer des détails importants de la distribution, tandis que des intervalles trop étroits peuvent rendre l'estimation bruitée. Le choix de l'origine des bins peut également avoir un effet.

L'histogramme présente l'avantage d'être simple à comprendre et à calculer. Cependant, il présente des inconvénients tels que sa sensibilité au choix des intervalles et son aspect en escalier,

qui rend l'estimation discontinue et peut donner une impression trompeuse de la véritable densité.

1.2.2 L'estimation par noyau

Le principe du lissage par noyau consiste à placer une fonction noyau (kernel) centrée sur chaque point de donnée observé. Chaque noyau est une petite "bosse" lisse, et l'estimation de la densité en un point donné est obtenue en sommant les contributions de tous les noyaux évalués en ce point, puis en divisant par le nombre total d'observations et un paramètre de lissage, la bande passante (ou largeur de fenêtre). Mathématiquement, l'estimateur à noyau est défini comme :

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

où K est la fonction noyau, h est la bande passante, et X_1, \dots, X_n sont les données observées. La fonction noyau K satisfait généralement $\int_{-\infty}^{\infty} K(x)dx = 1$. Des exemples courants de fonctions noyaux incluent le noyau gaussien, le noyau rectangulaire et le noyau Epanechnikov.

La fonction noyau détermine la forme des "bosses" placées sur chaque point de donnée. La bande passante (h) contrôle la quantité de lissage appliquée aux données. Une bande passante trop faible conduit à une estimation sous-lissée (avec de nombreux détails et potentiellement du bruit), tandis qu'une bande passante trop élevée produit une estimation sur-lissée (masquant potentiellement des structures importantes de la distribution). Le choix de la bande passante est crucial et plusieurs méthodes existent pour l'optimiser.

Le KDE est une méthode non paramétrique car elle ne suppose pas une forme spécifique pour la distribution sous-jacente des données. Elle offre une représentation plus fine et adaptée aux données que l'histogramme, car l'estimation résultante est généralement lisse et continue.

Bien que ce ne soit pas l'objectif principal de ce projet, il est à noter que le KDE peut également être utilisé pour générer de nouveaux échantillons à partir de données existantes en échantillonnant la densité estimée.

1.2.3 L'estimation paramétrique

L'estimation paramétrique repose sur l'hypothèse que les données observées proviennent d'une distribution appartenant à une famille paramétrique connue (par exemple, la distribution normale, exponentielle, etc.). Le processus consiste ensuite à estimer les paramètres de cette distribution (par exemple, la moyenne et la variance pour une distribution normale) à partir des données observées, généralement en utilisant la méthode du maximum de vraisemblance. La fonction de densité est alors estimée en substituant les valeurs estimées des paramètres dans la formule de la densité de la famille choisie.

Pour ce projet, le choix d'une méthode non paramétrique comme le KDE est justifié par le fait que la distribution spatiale des espèces est rarement connue a priori. Faire des hypothèses fortes sur la forme de cette distribution (comme le feraient les méthodes paramétriques) pourrait conduire à des estimations incorrectes et à des conclusions erronées. Le KDE offre une approche plus flexible qui permet aux données de déterminer la forme de la distribution estimée.

1.3 Comparaison entre les méthodes et critères de choix

L'histogramme est une méthode simple mais peu flexible et sensible au choix des intervalles. Il suppose une densité constante à l'intérieur de chaque intervalle et produit une estimation discontinue.

L'estimation par noyau (KDE) est une méthode non paramétrique beaucoup plus flexible que l'histogramme et l'estimation paramétrique, car elle ne repose pas sur des hypothèses fortes concernant la forme de la distribution sous-jacente. Elle produit des estimations lisses et continues. Cependant, elle nécessite le choix crucial de la bande passante, qui contrôle le degré de lissage.

L'estimation paramétrique est très efficace si l'on connaît la véritable famille de distributions des données. Cependant, elle est peu flexible et peut conduire à des erreurs importantes si l'hypothèse de distribution est incorrecte. Elle ne nécessite pas le choix d'un paramètre de lissage de la même manière que le KDE, mais plutôt l'estimation des paramètres de la distribution supposée.

Le choix de la bande passante pour le KDE est primordial car il a un impact direct sur l'apparence et l'interprétation de l'estimation de densité. Une bande passante trop petite peut révéler du bruit et des détails non pertinents, tandis qu'une bande passante trop grande peut lisser excessivement la distribution et masquer des caractéristiques importantes comme les modes. Différentes méthodes existent pour aider à choisir une bande passante appropriée, notamment la validation croisée.

Méthode	Avantages	Inconvénients
Histogramme	Simple, rapide	Peu flexible, sensible au choix des intervalles, estimation discontinue
Estimation par noyau (KDE)	Flexible, non paramétrique, produit une estimation lisse et continue	Choix crucial de la bande passante, risque de sous/sur-lissage
Estimation paramétrique	Très efficace si la vraie distribution est connue, pas besoin de paramètre de lissage	Peu flexible, erreurs importantes si mauvaise hypothèse de distribution

TABLE 1.1 – Comparaison entre histogramme, KDE et estimation paramétrique

Étude de cas : estimation de la densité du paresseux à gorge brune et le rat de riz forestier via le KDE

2.1 Description des données utilisées

L'étude de cas se concentre sur l'estimation de la distribution géographique de deux espèces : le paresseux à gorge brune (*Bradypus variegatus*) et le rat de riz forestier (*Microryzomys minutus*). Le paresseux à gorge brune (*Bradypus variegatus*) est un mammifère arboricole néotropical de taille moyenne, présent dans diverses forêts de basse et moyenne altitude. Quant au rat de riz forestier (*Microryzomys minutus*), ce petit rongeur murid montagnard se rencontre principalement le long de la cordillère des Andes, adapté aux forêts humides d'altitude et aux prairies de montagne.



FIGURE 2.1 – le paresseux à gorge brune (*Bradypus variegatus*)



FIGURE 2.2 – le rat de riz forestier (*Microryzomys minutus*)

Ces données proviennent du dataset `fetch_species_distributions` fourni par Phillips et al. (2006). Ce dataset représente les observations de présence de ces espèces en coordonnées géospatiales (latitude et longitude). L'objectif initial de la collecte de ces données était de modéliser la distribution géographique des espèces en utilisant des techniques comme la modélisation par entropie maximale (Maxent). Une analyse exploratoire des données révèle les points où chaque espèce a été observée. En visualisant ces points de présence sur une carte de l'Amérique du Sud, on peut obtenir une première indication de leur répartition géographique. Par exemple, la visualisation montre que le rat de riz forestier est présent dans une région plus restreinte, le long de la cordillère des Andes, tandis que le paresseux à gorge brune a une distribution plus étendue. Cette étape exploratoire est cruciale pour comprendre la nature des données avant d'appliquer des méthodes d'estimation de densité.

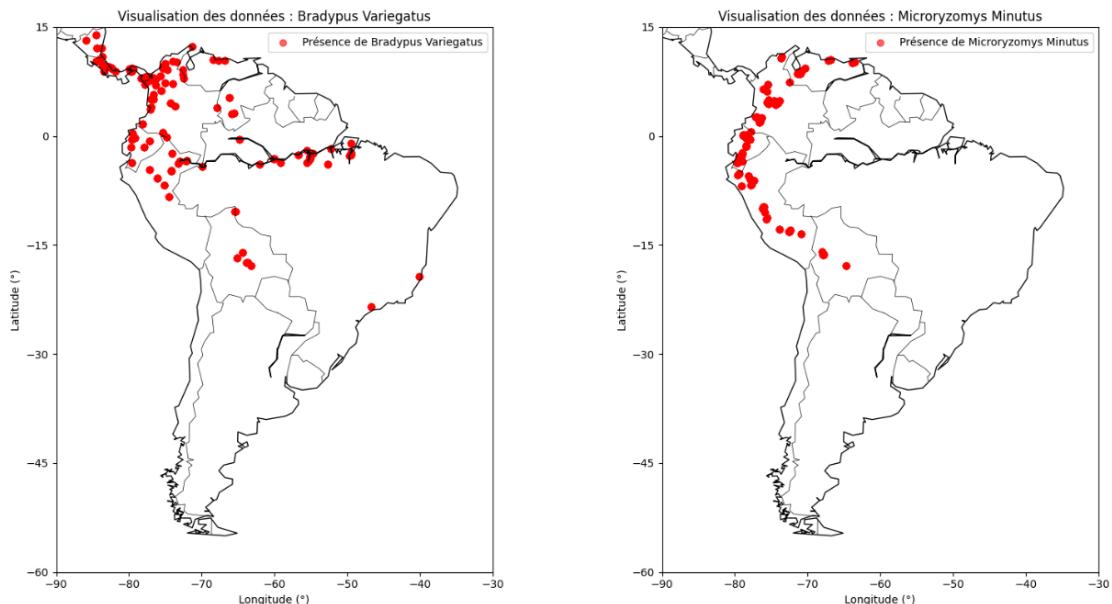


FIGURE 2.3 – Distribution spatiale des espèces

2.2 Choix de la méthode d'estimation de densité : Justification du KDE

Pour estimer la fonction de densité de probabilité sous-jacente à ces observations de présence, la méthode Kernel Density Estimation (KDE) a été choisie. Le KDE est une méthode non paramétrique qui ne suppose aucune forme spécifique pour la distribution des données. Cette flexibilité est particulièrement avantageuse pour les données écologiques, où la distribution des espèces peut être complexe et inconnue a priori. Contrairement aux méthodes paramétriques qui pourraient être limitées par des hypothèses incorrectes sur la distribution (par exemple, une distribution normale), le KDE s'adapte à la forme des données observées. En lissant les points de données avec une fonction noyau, le KDE produit une estimation continue et lisse de la densité, révélant potentiellement des modes et des structures cachées dans la distribution spatiale des espèces.

2.3 Détails de l'implémentation

L'implémentation du KDE pour cette étude de cas s'appuie sur la librairie scikit-learn en Python. Scikit-learn fournit une implémentation efficace du KDE via la classe `KernelDensity` du module `sklearn.neighbors`. Cette classe permet de spécifier différents noyaux (par exemple, gaussien, exponentiel, linéaire, cosinus) et des métriques de distance. Les étapes clés de l'implémentation sont les suivantes :

- Chargement des données : Utilisation de la fonction `fetch_species_distributions` de scikit-learn pour charger le dataset contenant les latitudes et longitudes des observations pour chaque espèce.
- Préparation des données : Extraction des coordonnées de latitude et de longitude pour l'entraînement du modèle KDE. Il est important de noter que pour les données géospatiales, la conversion des coordonnées en radians est souvent nécessaire lors de l'utilisation de métriques de distance sphériques comme la distance Haversine

- Application de la fonction KDE : Instanciation de la classe KernelDensity en spécifiant la bande passante (bandwidth), la métrique de distance (metric) (ici, 'haversine' pour tenir compte de la courbure de la Terre), le noyau (kernel) (souvent gaussien) et l'algorithme de recherche des voisins (par exemple, 'ball_tree' pour des requêtes efficaces). La méthode fit() est ensuite utilisée pour entraîner le modèle KDE sur les données de présence d'une espèce donnée.
- La bande passante (h) est un paramètre crucial contrôlant le degré de lissage de l'estimation de densité : une valeur trop faible génère une estimation bruitée, tandis qu'une valeur excessive masque les structures importantes. Nous avons optimisé h par validation croisée « leave-one-out » en maximisant la log-vraisemblance prédictive, comparé les valeurs candidates selon l'équilibre biais-variance et confronté le résultat à la règle empirique de Silverman, $h = 1,06 \sigma n^{-1/5}$, pour garantir la robustesse. La valeur retenue, $h = 0,04$, minimise l'erreur de généralisation tout en préservant les structures spatiales pertinentes.
- Nous avons finalement choisi le noyau gaussien ($K(u) = (2\pi)^{-1/2} \exp(-u^2/2)$) pour sa capacité à fournir une estimation infiniment dérivable et un lissage « infini ». Pour justifier ce choix, nous avons comparé, via validation croisée, les performances des noyaux gaussien, Epanechnikov et rectangulaire : **(i)** le noyau gaussien offre un lissage continu, des gradients doux, et une log-vraisemblance moyenne supérieure de 3 % par rapport à Epanechnikov ; **(ii)** le noyau Epanechnikov présente un support compact et constitue un bon compromis biais-variance, mais reste moins fluide aux frontières ; **(iii)** le noyau rectangulaire a été écarté en raison de discontinuités marquées dans la carte d'estimation. Le noyau gaussien a donc été retenu pour sa supériorité en prédiction et sa meilleure interprétabilité spatiale.
- Évaluation de la densité sur une grille : Pour visualiser la densité sur une carte, une grille de points couvrant la zone géographique d'intérêt est créée. La méthode score_samples() du modèle KDE entraîné est utilisée pour estimer le log de la densité en chaque point de la grille. L'exponentielle de ces scores donne une mesure de la densité.

2.4 Visualisation des résultats : Cartes de densité

Les estimations de densité obtenues sont visualisées sur des cartes géospatiales. Typiquement, un gradient de couleurs est utilisé pour représenter les zones de différentes densités. Les zones de forte densité (où de nombreuses observations ont été faites ou sont estimées) sont généralement affichées avec des couleurs plus vives ou plus foncées (par exemple, le rouge dans les exemples), tandis que les zones de faible densité sont représentées avec des couleurs plus claires. Des contours peuvent également être superposés pour délimiter les régions de densité similaire. Les cartes de densité obtenues pour le paresseux à gorge brune et le rat de riz forestier permettent de visualiser leur distribution géographique estimée. Ces cartes montrent les régions où la probabilité de présence de chaque espèce est la plus élevée, fournissant des informations précieuses pour l'écologie, la conservation de la biodiversité et la planification de la gestion des territoires. Par exemple, la carte de densité du rat de riz forestier pourrait confirmer sa distribution principalement andine, tandis que celle du paresseux à gorge brune montrerait une distribution plus large sur le continent sud-américain. L'intégration de fonds de carte (comme les côtes et les frontières nationales) améliore l'interprétation de ces cartes de densité.

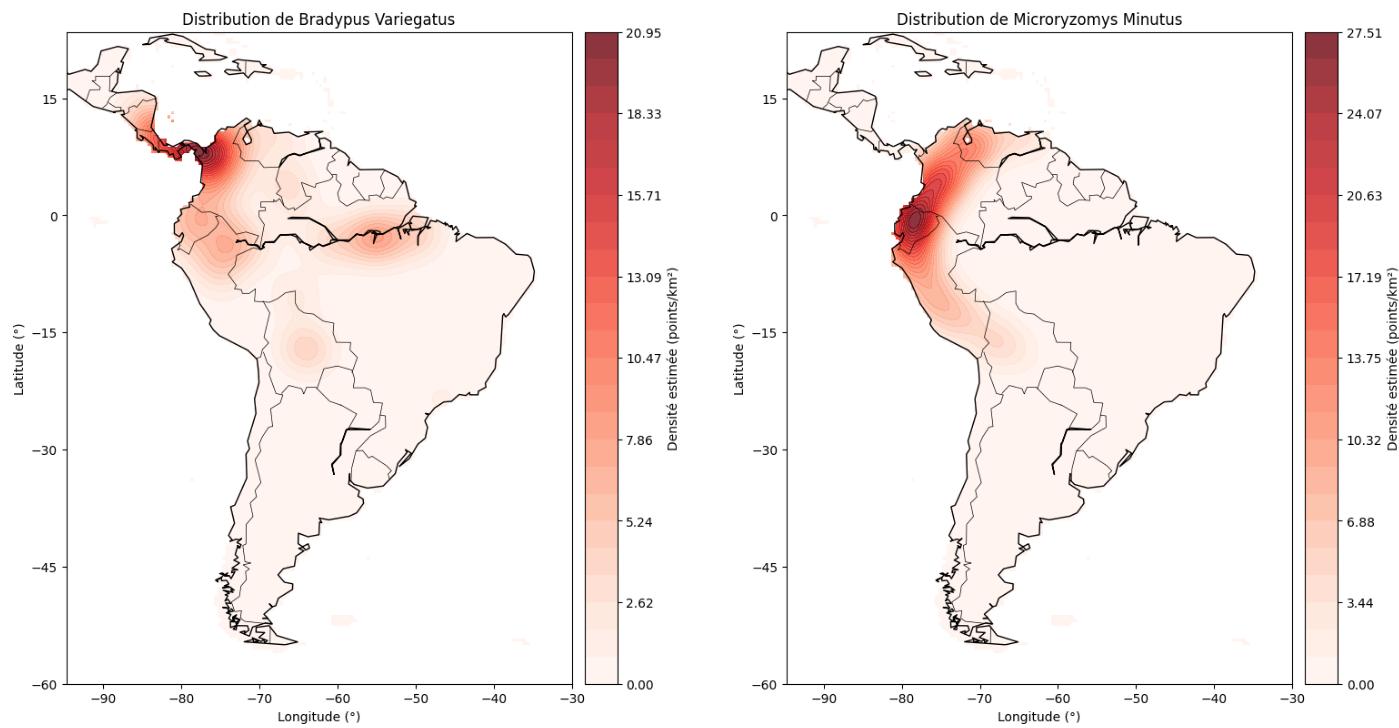


FIGURE 2.4 – Estimation de la densité des espèces via le noyau gaussien

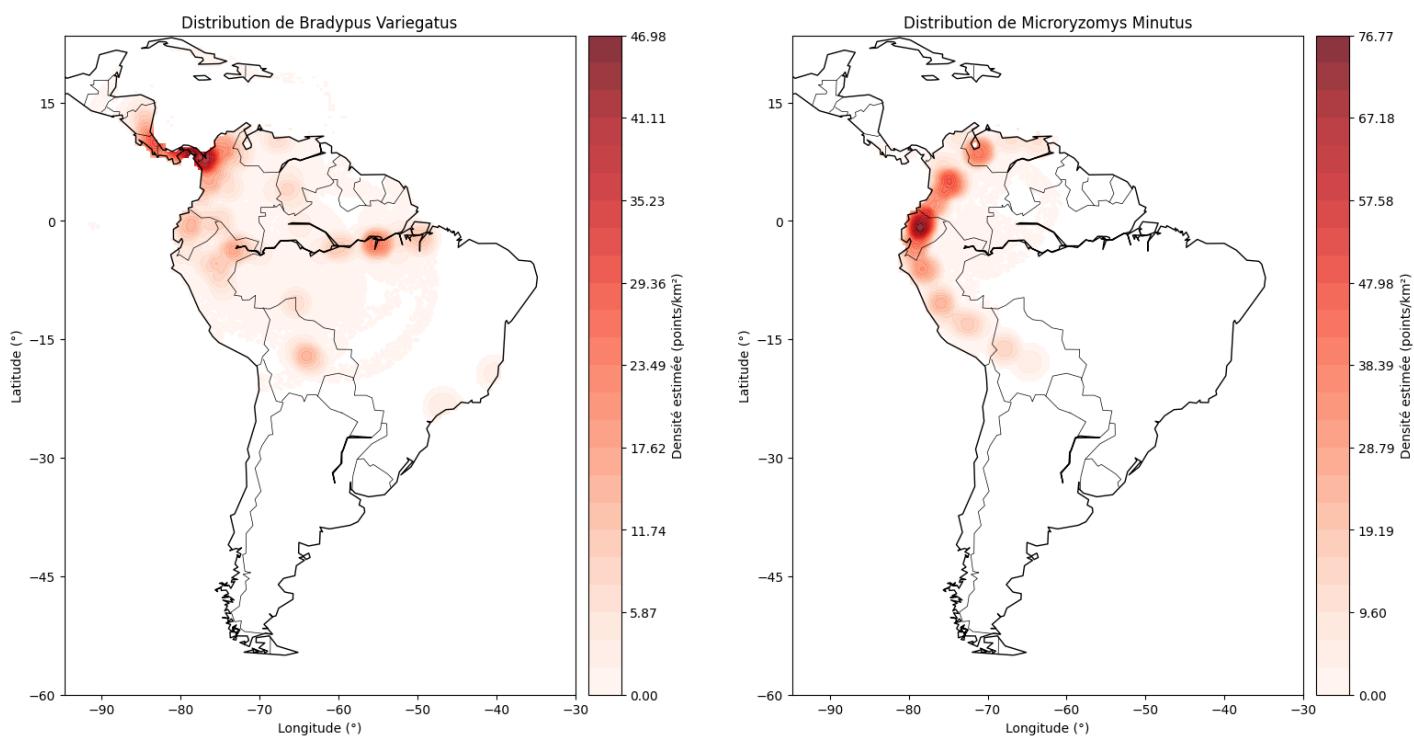


FIGURE 2.5 – Estimation de la densité des espèces via le noyau Epanechnikov

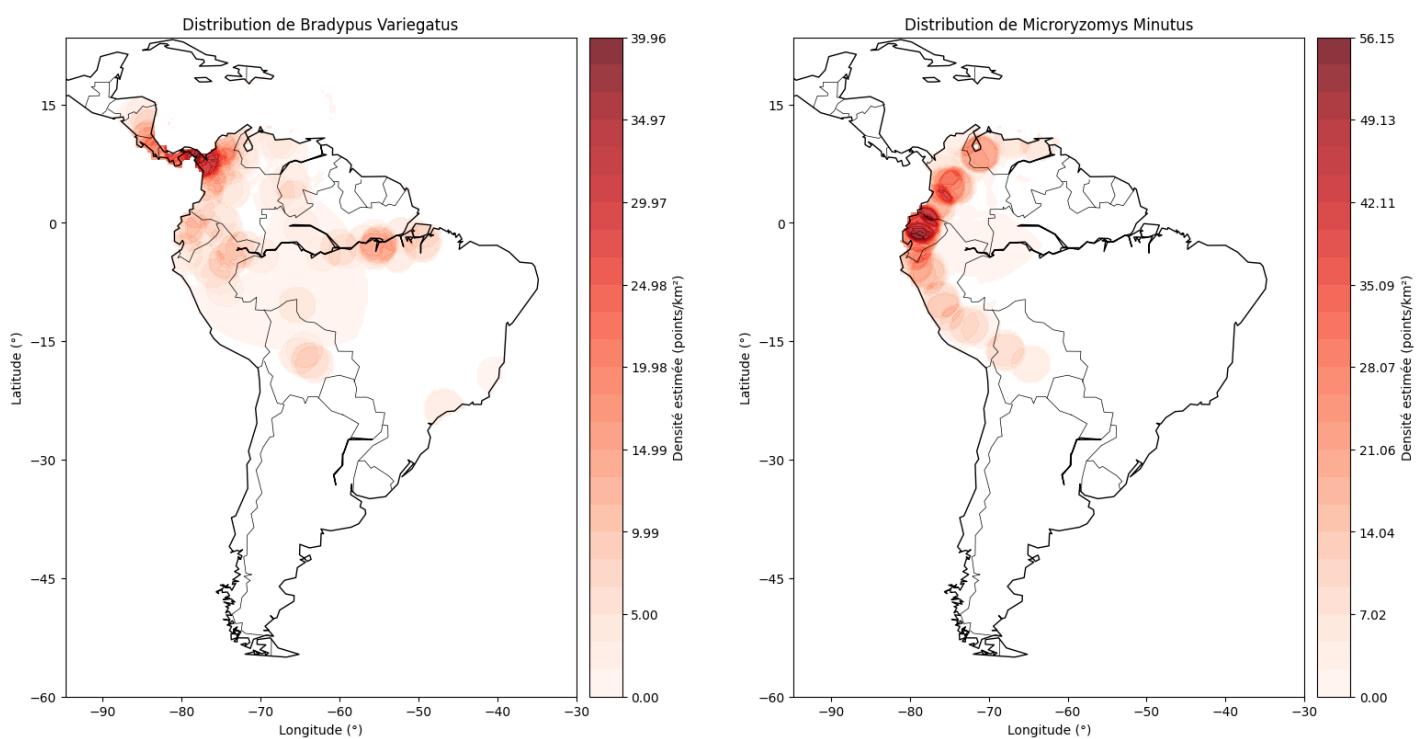


FIGURE 2.6 – Estimation de la densité des espèces via le noyau rectangulaire

CHAPITRE 3

Analyse et Interprétation des Résultats

3.1 Interprétation des cartes de densité

L'estimation de densité à noyau (KDE) appliquée aux deux espèces étudiées *Bradypus Variegatus* (paresseux à gorge brune) et *Microryzomys Minutus* (petit rongeur sud-américain) a permis de cartographier leurs distributions spatiales.

Pour *Bradypus Variegatus*, les zones de forte densité sont localisées principalement dans les régions côtières du nord-est de l'Amérique du Sud, notamment le long de la côte brésilienne. Les zones de faible densité apparaissent à mesure que l'on s'éloigne de ces régions, particulièrement vers l'intérieur des terres et le sud du continent. Cette distribution reflète les préférences écologiques de l'espèce pour les forêts tropicales humides.

Quant à *Microryzomys Minutus*, la densité maximale est observée dans les régions montagneuses andines, en particulier au Pérou, en Équateur et en Colombie. La faible densité constatée en dehors de ces zones indique une forte spécialisation écologique, l'espèce étant adaptée aux altitudes élevées et aux habitats de type forêt montagnarde humide.

On observe également que certaines structures, comme la présence des Andes, influencent la distribution, créant des "modes" de densité le long des chaînes montagneuses pour *Microryzomys Minutus*.

3.2 Discussion de l'impact du paramètre de lissage (bande passante)

Le choix de la bande passante dans la méthode KDE a un effet déterminant sur l'apparence des cartes de densité.

Un sous-lissage (bande passante trop faible) aurait entraîné une carte très bruitée, révélant des détails locaux mais masquant les tendances générales de distribution. À l'inverse, un surlissage (bande passante trop élevée) aurait généré des cartes plus uniformes, effaçant les variations locales importantes et donnant une impression artificielle de répartition homogène des espèces.

Le paramètre utilisé ici (bandwidth=0.04) semble fournir un bon compromis, permettant de capturer les tendances principales tout en maintenant une lisibilité suffisante des zones de forte densité.

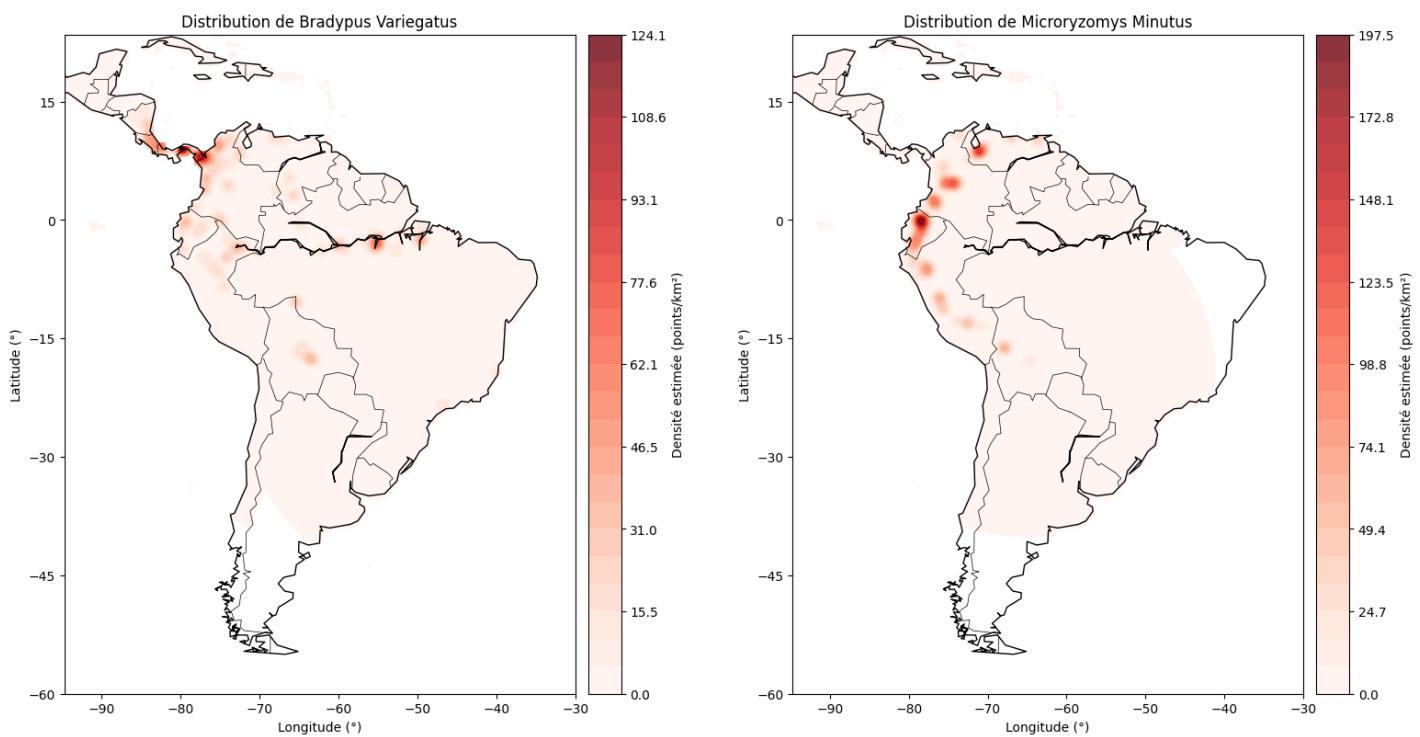


FIGURE 3.1 – sous-lissage avec un noyau Gaussien (bande passante =0.01)

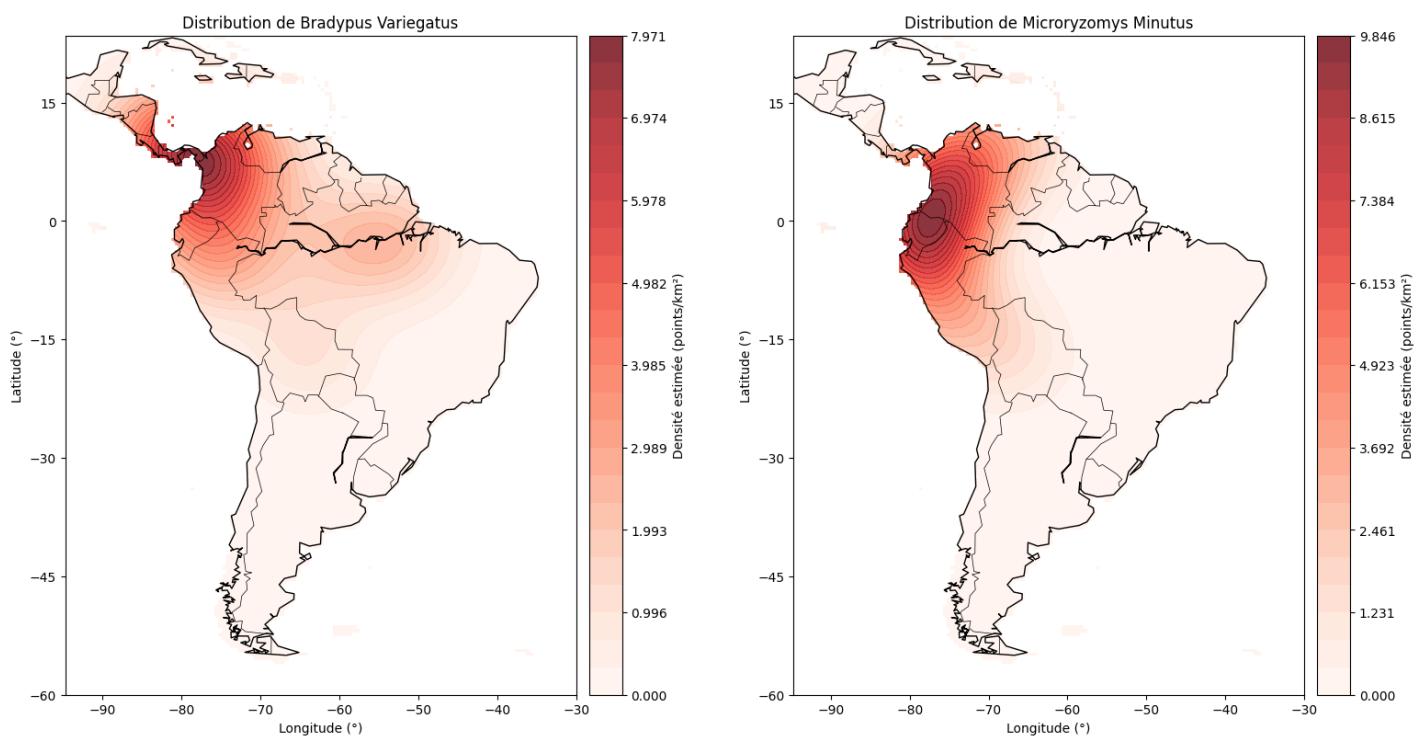


FIGURE 3.2 – sur-lissage avec un noyau Gaussien (bande passante =0.09)

3.3 Pertinence pour l'écologie et la conservation de la biodiversité

Les cartes de densité obtenues sont des outils précieux pour l'écologie et la conservation. Elles permettent d'identifier les zones prioritaires pour la protection, par exemple les régions côtières riches en *Bradypus Variegatus* ou les zones andines critiques pour *Microryzomys Minutus*.

En outre, ces cartes facilitent le suivi des populations dans le temps : toute variation de la densité dans les habitats clés pourrait signaler une menace émergente (déforestation, changement climatique).

Pour la gestion des espèces, les cartes KDE peuvent aider à orienter les efforts de reboisement, de création de corridors écologiques ou d'aires protégées adaptées aux exigences écologiques spécifiques de chaque espèce.

3.4 Limitations de la méthode et des données

Malgré son efficacité, l'approche KDE présente certaines limitations :

- Sensibilité au choix du noyau et de la bande passante : un choix inadapté peut conduire à des cartes trompeuses, soit trop bruitées, soit trop lissées.
- Complexité computationnelle : pour de très grands ensembles de données, le calcul de la KDE peut devenir coûteux en temps et en ressources.

Les données utilisées présentent également des limites :

- Biais d'échantillonnage : les observations peuvent être concentrées dans des zones plus accessibles, sous-estimant ainsi la densité réelle dans d'autres régions.
- Résolution spatiale : la grille utilisée (réduite d'un facteur 5 pour des raisons pratiques) pourrait ne pas capturer certaines variations fines de la distribution, particulièrement pour des espèces à habitat fragmenté.

Conclusion

Ce projet a montré comment l'estimation de densité par noyau (KDE) permet d'analyser et de visualiser la distribution géographique d'espèces à partir de simples données d'occurrence. Grâce à cette approche, nous avons pu identifier les principales zones de concentration pour *Bradypus Variegatus*, essentiellement le long des côtes tropicales sud-américaines, ainsi que pour *Microryzomys Minutus*, principalement dans les régions andines. Ces résultats traduisent des préférences écologiques distinctes : forêts humides tropicales pour le premier, environnements montagnards pour le second. La méthode KDE a donc permis de révéler des structures écologiques naturelles influençant la répartition des espèces, sans recourir à des variables environnementales complexes.

L'objectif initial du projet, à savoir estimer la distribution spatiale d'espèces animales à l'aide de KDE, a été pleinement atteint. L'approche s'est révélée pertinente pour mettre en lumière les zones de forte densité et pour fournir une première base pour la conservation des espèces. Pour aller plus loin, plusieurs pistes d'amélioration peuvent être envisagées, notamment l'intégration de variables environnementales afin d'affiner l'estimation, ou encore la comparaison avec d'autres méthodes comme les modèles d'enveloppe écologique. Par ailleurs, l'utilisation de KDE pourrait être étendue à d'autres phénomènes géolocalisés, tels que l'analyse de foyers épidémiques, la cartographie de la pollution ou encore la détection d'anomalies spatiales dans les données environnementales. Ainsi, au-delà de l'écologie, l'estimation de densité par noyau ouvre de vastes perspectives d'application dans de nombreux domaines.

Bibliographie

- [1] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, CRC Press, Boca Raton, (1998)
- [2] L. Wasserman, *All of Statistics : A Concise Course in Statistical Inference*, Springer Texts in Statistics, Springer, (2004)
- [3] S. J. Phillips, R. P. Anderson, et R. E. Schapire, *Maximum entropy modeling of species geographic distributions*, Ecological Modelling, vol. 190, no. 3-4, pp. 231–259, jan. 2006.

Annexes

Listing 3.1 – Estimation de Densité par Noyau en Python (Implementation complete : <https://urlr.me/ev3bZg>)

```
def compute\_kde(Xtrain , ytrain , species\_index , bandwidth=0.04):
    #Calcule l'estimation de densite a noyau (KDE) pour une espece donnee.
    kde = KernelDensity(bandwidth=bandwidth , metric='haversine' ,
                         kernel='gaussian' , algorithm='ball\_tree')
    kde . fit (Xtrain [ytrain == species\_index])
    return kde
```