



Rapport du projet de fin d'études sur l'analyse de permis de construire Projet de fin d'études

BARBOSA Mathias, LASGLEIZES David, THIRÉ Mael

ING3 IA1
2023 / 2024

Table des matières

1	Introduction et Motivation	2
2	État de l'art	3
2.1	Bunching : Un Phénomène Économique Étudié	3
2.2	À l'échelle de l'Union Européenne	4
2.3	Aux États-Unis	5
2.4	Conclusion	7
3	Traitement et exploration des données	9
3.1	Importation des données	9
3.2	Nettoyage des données : Uniformisation des codes de département	9
3.3	Exploration des données	9
3.4	Visualisation des données	9
4	Étude de trois variables clés	11
4.1	Hauteur des Bâtiments	11
4.2	Coefficient de Prise (Occupation) au Sol	12
4.3	Superficie du Terrain	12
5	Algorithme de détection bunching de données	14
5.1	Traitement des Données et Premières Visualisations	14
5.2	Variables clés étudiées	15
6	Détection de Bunching dans les Permis de Construire : Analyse des Algorithmes	16
6.1	Introduction	16
6.2	Méthode 1 : Analyse des Gradients	16
6.3	Méthode 2 : Estimation de la Densité par Noyau (KDE)	17
6.4	Base de Données d'Entraînement et Vérification des Algorithmes	17
6.5	Résultats et Discussion	17
6.6	Conclusion	18
7	Entraînement de différents modèles	19
7.1	Entraînement d'un modèle de Régression Linéaire permettant de vérifier la dépendance de la somme des cos par rapport aux autres variables . . .	19
7.2	Entraînement d'un modèle de Forêt Aléatoire	19
7.3	Entraînement d'un modèle de Réseau de Neurones Convolutif (CNN) . .	20
7.4	Comparaison des modèles	21
7.5	Conclusion	23
8	Conclusion	24
1	Annexe	26

1 Introduction et Motivation

Dans le cadre de notre projet de fin d'études, nous nous proposons d'aborder une problématique d'importance cruciale : l'analyse des permis de construire de logements dans les communes françaises à travers la détection de bunching, soutenu par SciencePo. Plutôt que de simplement retracer l'évolution de la réglementation, notre objectif est d'utiliser l'intelligence artificielle pour comprendre les dynamiques et les tendances qui ont façonné les politiques d'urbanisme en France.

Nous sommes conscients que des réglementations telles que l'article 1 du Code de l'urbanisme, adopté en 2005, ont profondément influencé la planification et l'autorisation des projets de construction au fil du temps. Toutefois, nous nous engageons à explorer au-delà de ces textes de loi, en examinant les influences multiples émanant d'instances nationales, européennes et locales.

Ainsi, notre étude se concentrera non seulement sur les réglementations en vigueur, mais également sur les pratiques et les tendances observées dans les permis de construire délivrés. Nous chercherons à comprendre comment ces facteurs interagissent pour façonner le paysage réglementaire et urbanistique français.

En résumé, notre projet vise à fournir une analyse approfondie et contextualisée des permis de construire en France, en combinant les perspectives réglementaires, pratiques et technologiques pour éclairer les décisions futures en matière d'urbanisme.

2 État de l'art

2.1 Bunching : Un Phénomène Économique Étudié

Le phénomène du "bunching" est un concept clé étudié en économie, notamment en économie du travail et en économie publique. Il se produit lorsque les individus ou les entreprises ajustent leur comportement de manière non continue mais plutôt en se regroupant autour de certains points ou seuils dans une distribution. Ce regroupement crée des pics ou des "bunches" dans la distribution, d'où le terme "bunching".

Objectif Économique : Incitations et Taxes

D'un point de vue économique, le bunching est souvent associé à des incitations ou à des réglementations fiscales. Par exemple, dans le cadre des impôts, les individus peuvent ajuster leur revenu juste en dessous d'un seuil pour bénéficier de réductions d'impôts ou d'autres avantages. De même, les entreprises peuvent ajuster leur taille ou leur production pour éviter de franchir certains seuils réglementaires ou fiscaux qui pourraient entraîner des coûts supplémentaires.

Utilité dans l'Analyse de Permis de Construire

Dans le domaine de l'urbanisme et de l'aménagement du territoire, l'analyse des permis de construire peut bénéficier de la compréhension du bunching. Par exemple, les réglementations municipales peuvent imposer des restrictions ou des incitations en fonction de la taille ou du type de projet de construction. Les développeurs immobiliers pourraient ajuster la taille de leurs projets pour éviter de dépasser certains seuils réglementaires, tels que des quotas de densité, des exigences de stationnement ou des contraintes environnementales. En identifiant les points de bunching dans les données des permis de construire, les planificateurs urbains et les décideurs politiques peuvent mieux comprendre comment les acteurs du marché réagissent aux réglementations et comment ces réglementations pourraient être modifiées pour atteindre des objectifs politiques spécifiques tout en minimisant les distorsions du marché.

L'Analyse Approfondie du Bunching dans les Permis de Construire

L'identification et l'analyse du bunching dans les données des permis de construire peuvent fournir des informations précieuses pour les planificateurs urbains et les décideurs politiques. Voici quelques aspects clés à considérer :

Compréhension des Incitations Économiques

En examinant les schémas de bunching dans les données des permis de construire, les planificateurs urbains peuvent mieux comprendre les incitations économiques qui motivent les développeurs immobiliers. Par exemple, si un grand nombre de projets de construction se regroupent juste en dessous d'un seuil spécifique de densité de construction, cela pourrait indiquer que les développeurs cherchent à éviter des exigences plus strictes ou des coûts supplémentaires associés à des densités plus élevées.

Évaluation de l'efficacité des Réglementations

En identifiant les points de bunching, les décideurs politiques peuvent évaluer l'efficacité des réglementations existantes sur les permis de construire. Si les données montrent que de nombreux projets de construction se regroupent juste en dessous d'un certain seuil, cela pourrait indiquer que la réglementation actuelle a des effets indésirables sur le marché immobilier, tels que des barrières à l'entrée pour les petits développeurs ou des distorsions de la concurrence.

Conception de politiques plus efficaces

En comprenant les comportements de bunching, les décideurs politiques peuvent concevoir des politiques plus efficaces et mieux ciblées pour atteindre les objectifs d'aménagement urbain. Par exemple, au lieu d'imposer des seuils rigides, ils pourraient envisager des mécanismes plus flexibles qui encouragent la densification progressive ou qui prennent en compte les caractéristiques spécifiques des quartiers.

Prévision des Tendances Futures

En surveillant les schémas de bunching dans le temps, les planificateurs urbains peuvent également anticiper les tendances futures du marché immobilier et ajuster leurs politiques en conséquence. Par exemple, si les données révèlent un bunching croissant autour de certaines caractéristiques de construction écologique, cela pourrait indiquer une demande croissante pour des projets respectueux de l'environnement, ce qui pourrait influencer les décisions d'investissement et les politiques futures.

En résumé, l'analyse du bunching dans les données des permis de construire offre une perspective précieuse sur les dynamiques du marché immobilier et peut guider la formulation de politiques urbaines plus efficaces et mieux adaptées aux besoins des collectivités locales.

2.2 À l'échelle de l'Union Européenne

Le rapport intitulé "An analysis of building construction based on building permits statistics" [2] constitue une exploration approfondie des tendances au sein de l'industrie de la construction dans l'Union européenne, en se penchant particulièrement sur les indices des permis de construire. Ces indices jouent un rôle crucial en tant qu'indicateurs reflétant l'évolution de l'activité de construction, soulignant ainsi l'importance de statistiques détaillées et fiables pour évaluer les développements au sein de la zone euro, une nécessité cruciale pour la politique monétaire commune. Les points clés abordés dans ce rapport sont résumés ci-dessous :

Deux indices principaux sont dérivés des statistiques des permis de construire :

Indice du nombre de logements : Mesure le nombre de logements résidentiels pour lesquels les permis ont été délivrés.

Indice de la surface utile autorisée : Englobe tous les types de bâtiments, y compris les établissements non résidentiels.

Les tendances au cours des dix dernières années : Le rapport met en évidence que les indices des permis de construire pour l'EU-27 ont atteint leur apogée en 2006 et au début de 2007, suivis d'un déclin qui reflète l'impact de la récente crise économique et financière. Les facteurs de

ce déclin incluent des surcapacités antérieures dans le secteur de la construction, une confiance réduite des consommateurs et des entreprises, un financement difficile pendant la crise, ainsi que des coupes dans les dépenses publiques.

Évolution pendant le cycle d'affaires de la construction : Des analyses de données, notamment des taux de croissance annuels, révèlent des tendances fortement positives entre 2002 et 2006 pour l'indice du nombre de logements. Cependant, les taux négatifs post-2007 ont significativement réduit l'indice de moitié.

Différences spécifiques aux pays : L'impact négatif majeur observé en 2008 et 2009 varie considérablement entre les États membres, certains, tels que l'Allemagne et la Slovaquie, enregistrant des expériences contrastées avec des croissances pendant ces années difficiles.

Comparaison trimestrielle et analyses spécifiques par pays : Des séries plus courtes sur une base trimestrielle sont présentées, mettant en lumière les fluctuations plus immédiates dans les indices. Les analyses spécifiques soulignent que la diminution des permis de construire anticipe souvent un ralentissement de l'activité de construction.

Précautions dans l'interprétation des indices : Il est souligné que tous les permis de construire ne se concrétisent pas nécessairement en projets de construction réels, soit en raison de leur non-utilisation, soit en raison du délai entre l'octroi du permis et le début des travaux.

Ainsi, le rapport de l'Eurostat démontre la pertinence cruciale des statistiques des permis de construire en tant qu'indicateurs de la santé économique du marché de la construction dans l'Union Européenne. Les variations dans ces indices révèlent des informations vitales sur l'état du secteur, tant pour les décideurs politiques que pour les acteurs du marché.

2.3 Aux États-Unis

L'article intitulé "The Emergence of Exclusionary Zoning Across American Cities" par Tianfang Cui [3] propose une nouvelle approche pour estimer l'adoption et la restriction des contrôles de zonage sur les lots dans les banlieues américaines.

Les algorithmes développés permettent de cartographier les contrôles de lot minimum dans les juridictions locales, en fournissant ainsi la première série temporelle sur l'adoption et la restrictivité des contrôles de lot minimum, une réglementation clé pour les banlieues américaines. L'entraînement se fait en trois phases :

1. Algorithme de détection de taille de lot : Cet algorithme utilise un apprentissage non supervisé pour analyser les distributions de taille de lot. Il identifie les lotissements qui présentent une concentration inhabituelle de propriétés autour de certaines tailles de lots, ce qui indique la présence de réglementations en matière de taille de lot. L'algorithme génère une liste de "bandes de concentration" pour chaque juridiction traitée.

2. Algorithme de boucle interne qui commence par calculer un gradient. Ce gradient est déterminé en évaluant la variation de la distribution de la taille des lots μ autour d'un point

de temps spécifique τ , en prenant en compte les lots spécifiques sur des périodes de temps plus et moins T . Une fois ce gradient calculé, l'algorithme effectue une version statique, où la largeur variable μ n'est pas considérée, afin de fournir des données agrégées à la boucle externe. Cette approche permet à la boucle externe d'évaluer l'impact des valeurs générées par la boucle interne.

Concrètement, la boucle interne itère sur différents points de temps τ et calcule les gradients associés en évaluant comment la distribution de la taille des lots évolue autour de ces points. Une fois cette phase terminée, les résultats agrégés sont utilisés dans la boucle externe pour prendre des décisions globales sur l'adoption et la restrictivité des contrôles de taille de lot. Cette approche en deux étapes permet de séparer les calculs locaux de la boucle interne des décisions globales prises par la boucle externe, améliorant ainsi la capacité de l'algorithme à saisir les tendances et les modèles dans les données à grande échelle.

3. Validation de l'algorithme : Un ensemble de données est divisé aléatoirement en un ensemble d'entraînement et un ensemble de test pour éviter le sur-ajustement. L'algorithme est évalué en termes de biais et de réduction de la variance pour les prévisions des dates d'adoption des tailles de lot. Des tests hors échantillon sont également effectués pour évaluer les performances de l'algorithme.

4. Choix des variables des paramètres de réglage i.e. les variables de test de la boucle externe pour évaluer les résultats de la boucle interne :

- α , définit les valeurs seuils autour desquelles les lots sont regroupés. Exemple, si α est défini à 1000, alors les lots situés à moins de 1000 unités de distance les uns des autres seront regroupés ensemble.
- μ^{miss} ajuste la largeur de la région de regroupement en fonction de la position ϵ dans la distribution des tailles de lots. Par exemple, tester le regroupement à 30 000 par rapport à 30 500 pieds carrés aura tous deux une région de regroupement d'une largeur de $\times 2000$ sur le côté gauche de chaque binôme. En modifiant ϵ , il est possible de contrôler la sensibilité du regroupement et d'adapter l'analyse pour mieux comprendre la distribution des tailles de lots, en tenant compte de la masse manquante dans les données.
- M^L représente "La Densité minimale sur le bac nécessaire à la classification". C'est un seuil de densité qui doit être atteint dans un bac pour qu'il soit classifié ou considéré comme significatif. Ce critère aide à déterminer si un bac particulier représente une concentration significative de propriétés autour de certaines tailles de lot. Par exemple, si la densité minimale sur le bac nécessaire à la classification est fixée à 0,025, cela signifie qu'un bac doit avoir une densité d'au moins 0,025 propriétés par unité de mesure spécifique (par exemple, par mile carré) pour être considéré comme significatif.
- \underline{F} est le "Seuil de la FCD pour le classificateur de valeurs modales". Ce seuil détermine à quel point une densité cumulée doit être élevée pour qu'une valeur soit considérée comme modale. Exemple pour $\underline{F} = 0.25$, la valeur la plus fréquente (modale) dans un ensemble de données doit représenter au moins 25 % de la distribution cumulative pour être considérée comme telle.

- M^{GROWTH} est le "Seuil du facteur de croissance pour les violations de la pré-période". Ce seuil est comparé au résultat d'un gradient statique pour évaluer si une augmentation significative de la densité ou d'autres caractéristiques des lots a eu lieu avant l'adoption des réglementations.

Les résultats indiquent que les contrôles de lot minimum ont été largement adoptés dans les banlieues entre 1945 et 1970, une période qui a vu environ quatre millions d'Américains noirs quitter le Sud pour des opportunités économiques. L'étude montre également que les réponses réglementaires locales à la migration noire ont entraîné une restriction de la densité résidentielle pour au moins 830 000 unités en dehors du Sud.

En outre, l'article met en évidence l'impact de la migration noire sur les résultats liés aux permis de construire, indiquant que les gouvernements locaux ont utilisé des contrôles de zonage pour préserver des aménagements raciaux endogènes. Ces résultats fournissent des indications essentielles pour comprendre l'impact des politiques de zonage sur la démographie et la planification urbaine.

L'auteur a mis en place un algorithme pour identifier l'adoption et la restrictivité des tailles minimales de lot dans les banlieues des États-Unis. Cet algorithme itère sur un ensemble de données comprenant 86 millions de propriétés pour calculer le nombre de propriétés construites autour des tailles de lot de regroupement après l'adoption de la réglementation.

De plus, les algorithmes tiennent compte des différents "vintages" de logements pour classer les tailles de lot en tant que "lots de regroupement" pour une juridiction donnée. Les résultats montrent que 18% des logements construits entre 1925 et 2010 étaient construits autour des tailles de lot de regroupement. Ces résultats sont ensuite utilisés pour estimer l'impact de la migration noire sur les réglementations en matière de taille de lot.

L'auteur a mesuré ces résultats en utilisant des statistiques descriptives, des régressions de moindres carrés ordinaires et des modèles à variables instrumentales pour quantifier l'effet de la composition noire sur l'adoption et la restrictivité des tailles de lot. Les résultats sont présentés sous forme de tableaux et de graphiques détaillés montrant les effets de la migration noire sur les réglementations en matière de taille de lot.

2.4 Conclusion

En conclusion, notre état de l'art a mis en lumière l'importance cruciale de l'analyse des permis de construire dans la compréhension des dynamiques urbaines et des politiques d'urbanisme, à la fois à l'échelle européenne, américaine et, potentiellement, au sein des communes françaises.

Dans l'Union Européenne, les statistiques des permis de construire jouent un rôle vital en tant qu'indicateurs de la santé économique du marché de la construction, fournissant des informations précieuses pour les décideurs politiques et les acteurs du marché. Les tendances observées au cours des dernières décennies mettent en lumière l'impact significatif des événements économiques sur l'activité de construction.

Aux États-Unis, l'utilisation d'algorithmes avancés permet de cartographier et d'analyser les réglementations de zonage, offrant ainsi des perspectives sur l'impact des politiques urbaines sur la démographie et la planification urbaine. Les résultats de ces études soulignent l'importance de comprendre les interactions complexes entre les réglementations locales et les dynamiques sociales et économiques.

Quant aux communes françaises, bien que l'application de telles méthodes n'ait pas encore été réalisée, notre projet vise à combler cette lacune en développant un programme innovant pour retracer l'évolution des réglementations locales à partir des permis de construire. En combinant des techniques d'apprentissage automatique et des analyses spatiales, nous aspirons à fournir une référence précieuse pour l'analyse de l'urbanisme en France.

En résumé, notre projet s'inscrit dans une démarche multidisciplinaire visant à mieux comprendre les processus de planification urbaine et à fournir des outils analytiques et cartographiques pour soutenir la prise de décision en matière d'aménagement du territoire.

3 Traitement et exploration des données

Dans cette section, nous détaillons les étapes de traitement et d'exploration des données effectuées dans le cadre de notre étude sur les permis de construire.

3.1 Importation des données

Le processus d'importation des données commence par la définition d'un répertoire contenant les fichiers CSV (`repertoire_csv`). En utilisant la bibliothèque `os`, nous récupérons la liste des fichiers CSV dans ce répertoire. Ensuite, chaque fichier CSV est lu à l'aide de `pandas` et importé dans une table nommée `permis_construire` dans la base de données SQLite nouvellement créée. Cette étape permet de centraliser toutes les données dans une base de données unique.

3.2 Nettoyage des données : Uniformisation des codes de département

Une étape cruciale du processus de traitement des données consiste à nettoyer et à prétraiter les données afin de garantir leur intégrité, leur cohérence et leur fiabilité pour les analyses ultérieures. Le nettoyage des données vise à corriger les erreurs, à gérer les valeurs manquantes, à supprimer les doublons et à uniformiser les formats, entre autres tâches.

Dans notre étude sur les permis de construire, nous avons effectué une opération de nettoyage spécifique en uniformisant les codes de département (`DEP_COD_DEP`). Cette uniformisation a été réalisée à l'aide d'une requête SQL exécutée sur la base de données SQLite. Elle garantit une représentation cohérente des codes de département, facilitant ainsi les analyses géographiques et les comparaisons entre différentes régions.

La requête SQL utilise une logique conditionnelle pour s'assurer que tous les codes sont représentés de manière homogène. Elle vérifie la longueur de chaque code de département et ajoute des zéros au début si nécessaire, conformément au format requis. Cette approche assure la cohérence des données et évite les erreurs.

3.3 Exploration des données

Une fois les données importées, le code fourni effectue plusieurs requêtes SQL pour explorer les données. La première requête (`query_sample`) récupère les sources distinctes présentes dans la table `permis_construire`, ce qui permet de vérifier que tous les fichiers ont été correctement importés. Ensuite, une seconde requête (`query_sample`) est utilisée pour afficher un échantillon des données, montrant les dix premières lignes de la table `permis_construire`. Ces requêtes fournissent un aperçu initial des données et vérifient leur intégrité après l'importation.

3.4 Visualisation des données

Après l'exploration initiale des données, le code génère des visualisations graphiques à l'aide de `matplotlib`. Tout d'abord, un graphique à barres empilées est créé pour représenter le nombre de bâtiments par année et par département, offrant ainsi une vue d'ensemble des données. Ensuite, un deuxième graphique à barres est produit pour mettre en évidence le nombre de bâtiments par année dans la région Île-de-France, offrant ainsi une perspective plus spécifique sur une région particulière.

En résumé, le code fourni réalise un processus complet de traitement et d'exploration des données, allant de l'importation initiale des données à la visualisation des résultats. Chaque étape est cruciale pour comprendre la nature des données et identifier des tendances ou des motifs significatifs.

4 Étude de trois variables clés

Dans cette section, nous examinerons en détail l'importance de trois variables clés dans l'analyse des permis de construire : la hauteur des bâtiments (nombre d'étages), le coefficient de prise au sol et la superficie du terrain. Chacune de ces variables joue un rôle crucial dans la planification urbaine et la réglementation des constructions.

4.1 Hauteur des Bâtiments

La hauteur des bâtiments, exprimée en nombre d'étages, est un indicateur clé dans l'analyse des permis de construire. Cette variable influence l'esthétique du paysage urbain, la densité verticale et la qualité de vie des habitants.

Importance de la Hauteur des Bâtiments : La hauteur des bâtiments est un élément essentiel dans la planification urbaine, car elle détermine la densité de population, l'utilisation de l'espace et l'aménagement du territoire. Une augmentation du nombre d'étages peut permettre une utilisation plus efficace du sol et répondre à la demande croissante de logements dans les zones urbaines densément peuplées. Cependant, une hauteur excessive peut entraîner des problèmes d'ombre, de vue obstruée et de surpeuplement, ce qui nécessite une réglementation appropriée pour assurer un développement urbain équilibré et durable (voir Figure 3 pour une représentation visuelle).

Résumé des Observations : Sur un ensemble de permis de construire portant sur des logements collectifs, il a été observé que la majorité de ces permis déclarent des bâtiments de plain-pied. Cependant, une analyse comparative avec des données de cartographie telles que Google Maps révèle des incohérences significatives. Des bâtiments initialement déclarés comme étant sans étages se révèlent en réalité comporter plusieurs étages lorsqu'ils sont examinés sur le terrain virtuel. Cette divergence entre les données déclarées dans les permis de construire et la réalité observée sur le terrain soulève des questions sur l'exactitude des informations fournies et la fiabilité du processus d'approbation des permis (voir Figure 7 pour une représentation visuelle).

Cas Spécifique : Dans le cas spécifique de la ville de Gennevilliers, une analyse approfondie des permis de construire révèle une distribution inhabituelle du nombre d'étages déclarés pour les logements collectifs. Alors que la majorité des permis mentionnent des bâtiments de plain-pied, une comparaison avec des données cartographiques telles que Google Maps révèle une discordance significative. En effet, plusieurs bâtiments initialement déclarés comme étant sans étages se révèlent en réalité comporter plusieurs niveaux lorsqu'ils sont examinés sur le terrain virtuel.

Cette incohérence entre les données déclarées et la réalité observée soulève des interrogations quant à la fiabilité des informations fournies dans les permis de construire de la ville de Gennevilliers. Elle met en lumière les possibles lacunes dans le processus d'approbation des permis et souligne l'importance d'une vérification rigoureuse des données pour assurer la conformité réglementaire et la cohérence dans le développement urbain.

L'identification de telles divergences dans le cas de Gennevilliers souligne la nécessité d'une analyse approfondie des données déclarées dans les permis de construire. Une meilleure com-

préhension de ces écarts peut permettre d'améliorer les pratiques de régulation urbaine et de garantir une planification urbaine plus précise et conforme aux besoins réels des habitants (voir Figure 6 pour une représentation visuelle du problème).

4.2 Coefficient de Prise (Occupation) au Sol

Le coefficient de prise au sol (COS) est un autre indicateur crucial dans l'analyse des permis de construire, influençant la densité urbaine et l'utilisation du sol. Ce coefficient représente la proportion de la superficie au sol d'un terrain pouvant être occupée par la construction. Une valeur élevée du COS indique une utilisation intensive du terrain, tandis qu'une valeur basse peut signaler une utilisation plus dispersée ou moins dense.

Importance du Coefficient de Prise au Sol : Le COS est un élément essentiel dans la planification urbaine car il permet de contrôler la densité des constructions et d'optimiser l'utilisation des ressources foncières. Une utilisation efficace du terrain peut contribuer à réduire l'étalement urbain, à préserver les espaces naturels et agricoles, et à favoriser une mobilité plus durable en concentrant les activités et les habitations dans des zones plus denses et accessibles. Cependant, une interprétation erronée du COS, souvent liée à des déclarations incorrectes sur le nombre d'étages des bâtiments, peut compromettre la capacité des autorités à réguler efficacement le développement urbain et à atteindre les objectifs de durabilité.

Il est important de noter que le nombre d'étages déclaré dans les permis de construire peut parfois être erroné, ce qui influe directement sur le calcul du COS. Par exemple, si un bâtiment est déclaré comme étant de plain-pied mais qu'il comporte en réalité plusieurs étages, la surface au sol occupée par la construction sera sous-estimée, ce qui entraînera une surestimation du COS. Cette incohérence souligne la nécessité d'une vérification rigoureuse des données déclarées dans les permis de construire pour assurer une évaluation précise de la densité urbaine et de l'utilisation du sol. Des exemples illustrant ces incohérences peuvent être consultés dans l'annexe (voir Figures 4 et 7).

4.3 Superficie du Terrain

La superficie du terrain est un facteur déterminant dans la planification des projets immobiliers, influençant la taille des constructions et la densité résidentielle. Les réglementations sur la superficie minimale des terrains visent à garantir des normes de confort, de sécurité et d'accessibilité pour les occupants.

Dans le cadre de cette étude, l'analyse se concentre principalement sur la superficie des terrains par rapport aux projets de construction autorisés. L'objectif est d'évaluer si les terrains respectent les exigences réglementaires en termes de superficie minimale et comment cela influe sur la planification des projets immobiliers. Des exemples de cette analyse sont présentés dans l'annexe (voir Figure 5).

Importance de la Superficie du Terrain :

La superficie du terrain influence directement la capacité d'aménagement et la densité de construction d'une zone donnée. Un terrain plus vaste offre plus de possibilités en termes de conception architecturale et d'aménagement paysager, ce qui peut contribuer à la création d'environnements urbains plus attrayants et fonctionnels. De plus, une superficie adéquate

permet de mieux répondre aux besoins de la population en matière de logement, de services et d'infrastructures, favorisant ainsi un développement urbain équilibré et durable.

5 Algorithme de détection bunching de données

5.1 Traitement des Données et Premières Visualisations

Traitement des Données

Le processus de traitement des données débute par la concaténation des fichiers CSV contenant les informations sur les permis de construire. Ces fichiers sont localisés dans un répertoire spécifique et sont importés un par un. Ensuite, ils sont insérés dans une base de données SQLite nouvellement créée. Cette approche garantit que toutes les données sont consolidées en une seule source.

Vérification de la Qualité des Données

Pour assurer la qualité des données, deux étapes de vérification sont effectuées. Tout d'abord, une requête est exécutée pour vérifier les différentes sources des données, assurant ainsi que toutes les sources sont correctement intégrées. Ensuite, un échantillon de données est affiché pour s'assurer que l'importation s'est déroulée correctement et qu'il n'y a pas de problèmes majeurs de format ou de cohérence.

Uniformisation des Informations

Une fois les données importées, une uniformisation des informations est réalisée. Dans ce contexte, la colonne contenant les codes départementaux est normalisée pour garantir une cohérence et une manipulation aisée lors des analyses ultérieures. Cette uniformisation implique l'ajout de zéros en tête si nécessaire pour obtenir des codes départementaux sur trois chiffres.

Premières Visualisations

Les premières visualisations sont réalisées afin d'obtenir une compréhension initiale des données. Cela comprend une analyse du nombre de bâtiments par année et par département, ainsi qu'une analyse spécifique sur l'Île-de-France. Les résultats sont présentés sous forme de graphiques à barres empilées, permettant une visualisation claire des tendances au fil du temps et des variations entre les départements.

5.2 Variables clés étudiées

Dans le cadre de cette étude, trois variables clés ont été minutieusement examinées pour comprendre les tendances et les caractéristiques des constructions en Île-de-France :

Hauteur moyenne des bâtiments en Île-de-France

L'analyse de la hauteur moyenne des bâtiments en Île-de-France a commencé par l'extraction des données des permis de construire de la base de données. Ensuite, ces données ont été agrégées par année et par département, et la hauteur moyenne des bâtiments a été calculée pour chaque combinaison année-département. Ce calcul a impliqué la prise en compte de la hauteur déclarée dans les permis de construire. Par exemple, si un permis de construire déclarait une hauteur de 5 étages, cette valeur a été incluse dans le calcul de la hauteur moyenne pour l'année et le département correspondants. Les résultats de ces calculs ont été représentés graphiquement sous forme de graphique à barres.

Coefficient de prise au sol

Pour analyser le coefficient de prise au sol, les données des permis de construire ont été extraites de la base de données, en mettant l'accent sur les informations relatives à ce coefficient par année et par département. Ensuite, la somme spécifiée ($SHONANT + SHONCR - SHONDEM$) a été calculée pour chaque permis de construire, représentant ainsi la surface totale construite. Ces valeurs ont ensuite été moyennées pour chaque année et chaque département, fournissant ainsi une mesure moyenne du coefficient de prise au sol dans la région au fil du temps. Ces résultats ont été présentés graphiquement sous forme de graphique à barres pour une meilleure visualisation.

Superficie du terrain

L'examen de la superficie du terrain a impliqué le calcul de la superficie totale du terrain pour chaque permis de construire. Ces calculs ont été effectués à partir des données déclarées dans les permis, en se concentrant sur les informations relatives à la superficie du terrain par année et par département. Une fois la superficie totale du terrain calculée pour chaque permis, les données ont été agrégées par année et par département, et la superficie moyenne du terrain a été calculée. Ces résultats ont été représentés graphiquement sous forme de graphique à barres pour une visualisation facile.

Remise en question de la variable du nombre d'étages

La variable du nombre d'étages a été soumise à une vérification géospatiale approfondie pour évaluer sa validité. Après avoir extrait les données des permis de construire, une vérification géospatiale a été réalisée pour comparer les données déclarées avec la réalité sur le terrain. Cette vérification a permis d'identifier les adresses où le nombre d'étages déclaré était zéro. Ensuite, une enquête plus approfondie a été menée pour ces adresses afin de vérifier si elles avaient effectivement zéro étage ou si d'autres facteurs, tels que l'absence d'adresse ou des erreurs de déclaration, entraient en jeu. Cette analyse a souligné l'importance d'une collecte de données précise et fiable pour garantir la validité des analyses futures.

6 Détection de Bunching dans les Permis de Construire : Analyse des Algorithmes

6.1 Introduction

Dans cette section, nous présentons deux méthodes de détection de bunching appliquées à l'analyse des permis de construire. Le bunching est un phénomène où les observations se regroupent autour de certaines valeurs, souvent lié à des comportements stratégiques ou à des régulations spécifiques.

6.2 Méthode 1 : Analyse des Gradients

L'analyse des gradients est une méthode utilisée pour détecter le bunching en calculant les variations significatives dans la distribution des superficies des permis de construire entre deux périodes de temps.

Prétraitement des Données

Les données sont extraites de la base de données des permis de construire, filtrées par commune et par période. Ensuite, les superficies des permis de construire sont normalisées pour réduire les biais liés à la taille des constructions.

Calcul du Gradient

Le gradient des densités de permis de construire entre les deux périodes est calculé en utilisant des techniques de différenciation numérique. Soit $f(x)$ la fonction de densité des permis de construire à un certain point x , le gradient $\nabla f(x)$ est calculé en utilisant des méthodes telles que la différence finie centrée :

$$\nabla f(x) \approx \frac{f(x+h) - f(x-h)}{2h}$$

Où h est une petite valeur déterminée empiriquement pour définir la taille de l'intervalle autour de x pour le calcul du gradient.

Analyse Visuelle

Des visualisations telles que les histogrammes sont utilisées pour mieux comprendre la distribution des données. Ces visualisations permettent d'identifier visuellement les régions où le bunching est le plus probable en mettant en évidence les zones où le gradient est particulièrement élevé.

Identification du Bunching

En analysant les gradients calculés, les régions où le bunching est le plus probable sont identifiées. Les régions avec des gradients élevés indiquent une concentration anormale de permis de construire, ce qui suggère la présence de bunching.

6.3 Méthode 2 : Estimation de la Densité par Noyau (KDE)

L'estimation de la densité par noyau (KDE) est une méthode utilisée pour détecter le bunching en estimant la densité des données de permis de construire.

Estimation de la Densité

La méthode de KDE est utilisée pour estimer la densité des données de permis de construire. Elle consiste à placer un noyau autour de chaque observation et à calculer la densité en fonction du nombre de noyaux présents dans une région donnée. Une formule courante pour l'estimation de la densité est :

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

Où $\hat{f}(x)$ est la densité estimée à un point x , n est le nombre d'observations, h est la largeur de la fenêtre de noyau, et K est la fonction noyau.

Identification des Régions de Bunching

En analysant les pics de densité résultant de l'estimation de la densité par noyau, les régions où le bunching est le plus probable sont identifiées. Les pics de densité indiquent des concentrations anormales de permis de construire, suggérant la présence de bunching.

Comparaison avec les Données Précédentes

Les résultats de l'estimation de la densité sont comparés avec les données précédentes pour évaluer les changements dans la distribution des permis de construire. Les régions avec des pics de densité significativement plus élevés dans les données actuelles par rapport aux données précédentes sont considérées comme des régions où le bunching s'est produit.

6.4 Base de Données d'Entraînement et Vérification des Algorithmes

Nous détaillons ici la construction d'une base de données spécifiquement conçue pour entraîner et évaluer les performances des algorithmes de détection de bunching. Cette base de données comprend un ensemble représentatif de permis de construire avec des caractéristiques variées telles que la taille, l'emplacement et la période d'émission. L'objectif est d'ajuster les paramètres des algorithmes et de les tester sur des données diversifiées afin d'évaluer leur capacité à détecter efficacement le bunching.

6.5 Résultats et Discussion

Nous présentons les résultats obtenus lors de l'application des algorithmes de détection de bunching aux données réelles de permis de construire. Nous analysons les régions où le bunching est détecté et comparons les performances des deux méthodes présentées.

Discussion :**Analyse des Gradients :**— **Avantages :**

1. Simplicité d'Implémentation
2. Visualisation Intuitive
3. Sensibilité aux Changements Brusques

— **Inconvénients :**

1. Sensibilité aux Erreurs de Mesure
2. Dépendance aux Paramètres
3. Interprétation Subjective

Estimation de la Densité par Noyau (KDE) :— **Avantages :**

1. Adaptabilité
2. Estimation de la Densité
3. Faible Sensibilité aux Paramètres

— **Inconvénients :**

1. Complexité Computationnelle
2. Sensibilité à la Taille de l'Échantillon
3. Biais Potentiels

Cette analyse inclut également une évaluation des avantages et des limites de chaque approche, notamment en termes de sensibilité, de spécificité et de robustesse.

6.6 Conclusion

En conclusion, nous avons examiné les différentes méthodes de détection de bunching appliquées à l'analyse des permis de construire. Bien que ces approches aient montré leur efficacité dans la détection de comportements anormaux, elles présentent également des défis et des limitations. Nous soulignons l'importance de poursuivre les recherches pour améliorer la précision et l'efficacité de la détection de bunching, notamment en explorant de nouvelles techniques d'analyse de données et en intégrant des données supplémentaires pour une meilleure contextualisation des résultats.

7 Entraînement de différents modèles

7.1 Entraînement d'un modèle de Régression Linéaire permettant de vérifier la dépendance de la somme des cos par rapport aux autres variables

Nous débutons en chargeant les données à partir du fichier Excel et en sélectionnant les colonnes pertinentes, notamment '`c_coinsee`' et '`cos`'. Ensuite, nous préparons les données en convertissant les valeurs. Ensuite, nous calculons les statistiques regroupées par '`c_coinsee`', incluant la somme, le maximum, la moyenne, le type des superficies des permis de construire, ainsi que les déciles de la distribution. Ces informations s'avèrent cruciales.

Une fois les données préparées, nous les divisons en variables indépendantes (\mathbf{X}) et la variable dépendante (\mathbf{y}). Nous utilisons ensuite la fonction `train_test_split` pour séparer les données en ensembles d'entraînement et de test, avec une taille de test de 20%.

Par la suite, nous instancions et entraînons un modèle de régression linéaire à l'aide de la fonction `LinearRegression` de `scikit-learn`. Une fois le modèle entraîné, nous effectuons des prédictions sur l'ensemble de test.

Enfin, nous évaluons les performances du modèle en calculant l'erreur quadratique moyenne (MSE) et le coefficient de détermination (R^2). Le MSE représente la moyenne des carrés des erreurs entre les valeurs prédites et les valeurs réelles, tandis que le coefficient de détermination (R^2) mesure la proportion de la variance dans la variable dépendante qui est expliquée par la variable indépendante.

Les résultats de la régression linéaire sont les suivants :

- Erreur quadratique moyenne (MSE) : 5.766551851003644e-28
- Coefficient de détermination (R^2) : 1.0

Ces résultats indiquent que le modèle de régression linéaire explique la totalité de la variance des données, ce qui suggère que la variable somme de cos réelle est dépendante du reste.

7.2 Entraînement d'un modèle de Forêt Aléatoire

Nous utilisons un modèle de forêt aléatoire pour prédire si la valeur maximale de la superficie des permis de construire est supérieure à zéro, afin de classer les observations en deux catégories.

- Création d'une nouvelle variable cible binaire en fonction de la valeur maximale réelle des superficies des permis de construire. Cette variable est définie comme 1 si la valeur maximale est supérieure à zéro, sinon 0.
- Diviser les données en variables explicatives (\mathbf{X}) et la cible (\mathbf{y}). Les caractéristiques incluent les colonnes '`c_coinsee`', '`reel_cos_sum`', '`reel_cos_decile1`', '`reel_cos_decile2`', '`reel_cos_decile3`', '`reel_cos_decile4`', '`reel_cos_decile5`', '`reel_cos_decile6`', '`reel_cos_decile7`', '`reel_cos_decile8`', '`reel_cos_decile9`'.
- Diviser les données en ensembles d'entraînement et de test avec une taille de test de 20%.
- Initialiser le modèle de forêt aléatoire avec 100 arbres de décision et une graine aléatoire fixée à 420.
- Entraîner le modèle sur l'ensemble d'entraînement.
- Faire des prédictions sur l'ensemble de test.

- Évaluer les performances du modèle en calculant l'exactitude (accuracy) et en affichant le rapport de classification.

Les résultats de la forêt aléatoire sont les suivants :

- Accuracy : 0.96

Classification Report :

	precision	recall	f1-score	support
0	0.86	1.00	0.92	6
1	1.00	0.95	0.97	19
accuracy			0.96	25
macro avg	0.93	0.97	0.95	25
weighted avg	0.97	0.96	0.96	25

Ces résultats montrent que le modèle de forêt aléatoire a une précision globale de prédiction de 96%, avec des scores de précision, de rappel et de F1 très élevés pour les deux classes.

7.3 Entraînement d'un modèle de Réseau de Neurones Convolutif (CNN)

Enfin, nous utilisons un modèle de réseau de neurones convolutif (CNN) pour prédire si la valeur maximale de la superficie des permis de construire est supérieure à zéro. Voici les étapes détaillées :

- Les données sont préparées en normalisant les features et en appliquant l'oversampling pour équilibrer les classes.
- Les données sont divisées en ensembles d'entraînement et de test.
- Un modèle CNN est défini avec les couches suivantes :
 - Couche de convolution avec 32 filtres, une taille de noyau de 3x3 et une fonction d'activation ReLU.
 - Couche de normalisation en batch.
 - Couche de max pooling avec une taille de fenêtre de 2x2.
 - Couche de convolution avec 64 filtres, une taille de noyau de 3x3 et une fonction d'activation ReLU.
 - Nouvelle couche de normalisation en batch.
 - Nouvelle couche de max pooling avec une taille de fenêtre de 2x2.
 - Couche d'aplatissement (Flatten).
 - Couche dense avec 128 neurones et une fonction d'activation ReLU.
 - Couche de dropout avec un taux de 0.5.
 - Couche de sortie dense avec 1 neurone et une fonction d'activation sigmoïde.
- Le modèle est compilé avec l'optimiseur Adam et la perte binaire_crossentropy sur 100 epochs avec un batch size de 8.
- Le modèle est entraîné sur les données d'entraînement avec validation croisée.

- Des prédictions sont faites sur l'ensemble de test.
- Les performances du modèle sont évaluées en calculant la précision, le rappel, le score F1 et en affichant la matrice de confusion.

Les résultats du modèle CNN sont les suivants :

- Précision : 1.0
- Rappel : 1.0
- Score F1 : 1.0

Ces résultats indiquent une performance parfaite du modèle CNN dans la prédiction de la valeur maximale de la superficie des permis de construire. Une précision, un rappel et un score F1 de 1.0 signifient que le modèle prédit correctement toutes les instances de manière cohérente.

La matrice de confusion pour le modèle CNN est la suivante :

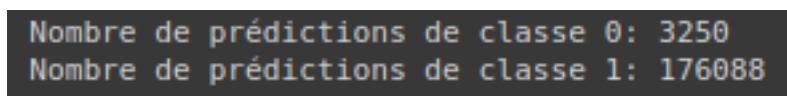
$$\text{Matrice de confusion : } \begin{bmatrix} 14 & 0 \\ 0 & 22 \end{bmatrix}$$

Dans cette matrice, les éléments diagonaux représentent les prédictions correctes, tandis que les éléments hors diagonale représentent les prédictions incorrectes. Dans ce cas, le modèle a correctement prédit 14 instances de la classe négative et 22 instances de la classe positive.

7.4 Comparaison des modèles

Les deux derniers modèles ont pour but de prédire sur des territoires locaux inconnus s'il y a une réglementation de coefficient de prise au sol maximum ou non.

Lorsque l'on compare les deux modèles précédents, on remarque beaucoup de différences. En effet, le dernier modèle semble prédire une très importante proportion de 1, ce qui peut être expliqué par le déséquilibre des classes. Or, même après rééquilibrage et entraînement d'un modèle sans overfitting, il continue à prédire beaucoup de 1 comparé au modèle de forêt aléatoire, comme le montre la figure suivante :



```
Nombre de prédictions de classe 0: 3250
Nombre de prédictions de classe 1: 176088
```

FIGURE 1 – Déséquilibre des prédictions

C'est pourquoi, le modèle de forêt aléatoire semble mieux convenir à notre problématique que le modèle neuronal. Ci-dessous un aperçu des prédictions des deux modèles :

FIGURE 2 – Quelques-unes des prédictions

7.5 Conclusion

Ainsi, on a pu prédire sur des territoires dont on n'a aucune information sur la réglementation, s'il y a une présence de règle de coefficient de prise au sol maximum ou non. Nos prédictions sont synthétisées dans un fichier csv. Nous nous sommes uniquement focalisés sur cette variable, car la validité des deux autres a été remise en question par le travail précédent.

8 Conclusion

En conclusion, notre projet de recherche a représenté une avancée significative dans la compréhension et l'analyse des permis de construire, mettant en lumière l'importance cruciale de cette démarche dans la planification urbaine et les politiques d'urbanisme. À travers une approche multidisciplinaire, nous avons exploré les dynamiques des permis de construire à l'échelle européenne, américaine et française, en utilisant des techniques avancées d'analyse de données et de modélisation.

Dans l'Union Européenne, nous avons souligné le rôle essentiel des statistiques des permis de construire en tant qu'indicateurs économiques clés, fournissant des informations précieuses pour les décideurs politiques et les acteurs du marché. Aux États-Unis, l'application d'algorithmes avancés a permis de mieux comprendre les réglementations de zonage et leur impact sur la démographie et la planification urbaine.

Notre projet vise spécifiquement à combler une lacune dans l'analyse des permis de construire au niveau des communes françaises, en développant un programme novateur pour retracer l'évolution des réglementations locales. En combinant des techniques d'apprentissage automatique et des analyses spatiales, nous avons cherché à fournir une référence précieuse pour l'urbanisme en France.

L'application des méthodes de détection de bunching à nos données réelles de permis de construire a permis d'identifier des comportements anormaux et d'évaluer les performances des différents algorithmes. Bien que ces approches aient montré leur efficacité, elles présentent également des défis et des limitations, soulignant ainsi la nécessité de poursuivre les recherches pour améliorer la précision et l'efficacité de la détection de bunching.

Enfin, notre travail s'est également concentré sur l'entraînement de modèles de prédiction, notamment en ce qui concerne les réglementations de coefficient de prise au sol maximum. Bien que notre étude se soit focalisée sur cette variable, elle ouvre la voie à de futures investigations pour explorer d'autres aspects des permis de construire et de la planification urbaine.

En somme, notre projet représente une contribution significative à la recherche en urbanisme, offrant des outils analytiques et des perspectives précieuses pour soutenir la prise de décision en matière d'aménagement du territoire. Nous espérons que nos travaux encourageront de nouvelles études dans ce domaine et contribueront à une planification urbaine plus efficace et éclairée.

Bibliographie

- [1] L.421 n°1-9, 9 décembre 2005, Code de l'urbanisme <https://www.legifrance.gouv.fr/codes/id/LEGISCTA0000006158675>
- [2] An analysis of building construction based on building permits statistics, 2010, Eurostat <https://op.europa.eu/en/publication-detail/-/publication/980b5f94-4ef9-48fa-8294-1e76f577d393/language-en>
- [3] The Emergence of Exclusionary Zoning Across American Cities, 2023, Tianfang Cui https://www.tom-cui.com/assets/pdfs/LotsEZ_Latest.pdf
- [4] Predicting Fine Particulate Matter (PM2.5) in the Greater London Area : An Ensemble Approach using Machine Learning Methods https://cran.r-project.org/web/packages/bunchr/vignettes/bunching_with_bunchr.html
- [5] Évaluation d'impact de la bascule du Crédit d'impôt pour la compétitivité et l'emploi (CICE) en allègement de cotisations employeur, Antoine Bozio, Sophie Cottet, Clément Malgouyres <https://shs.hal.science/halshs-03828744/document>

1 Annexe

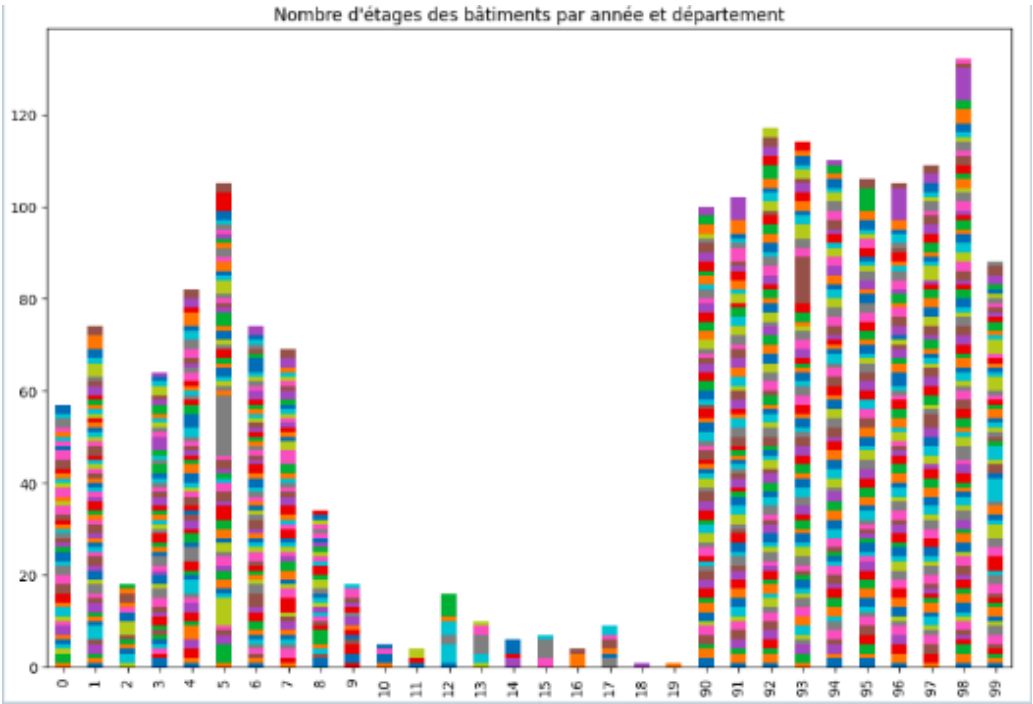


FIGURE 3

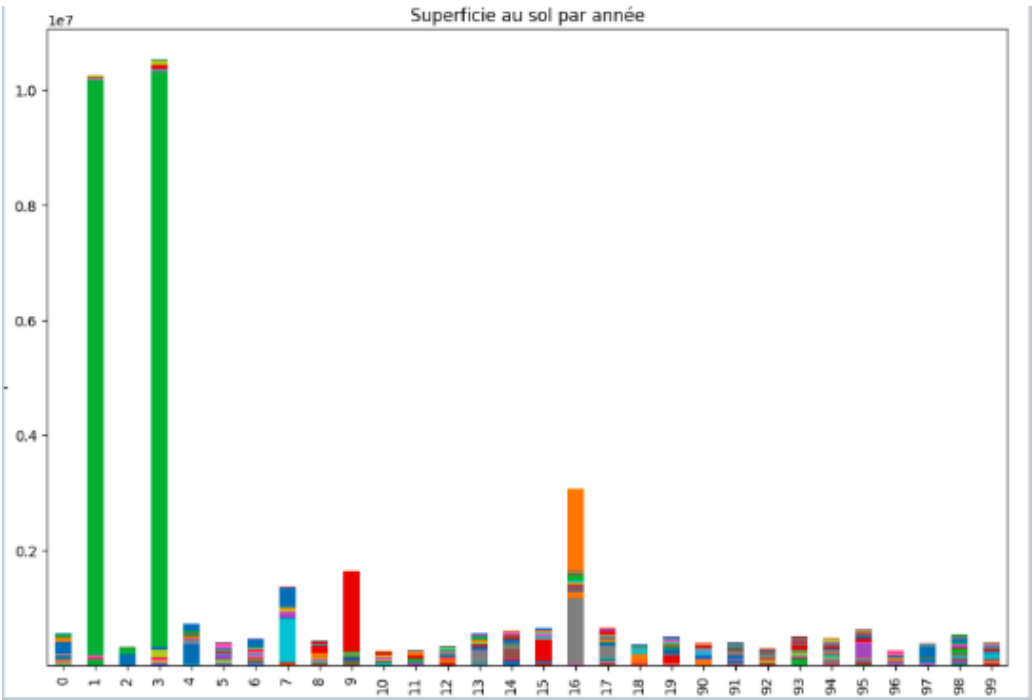


FIGURE 4

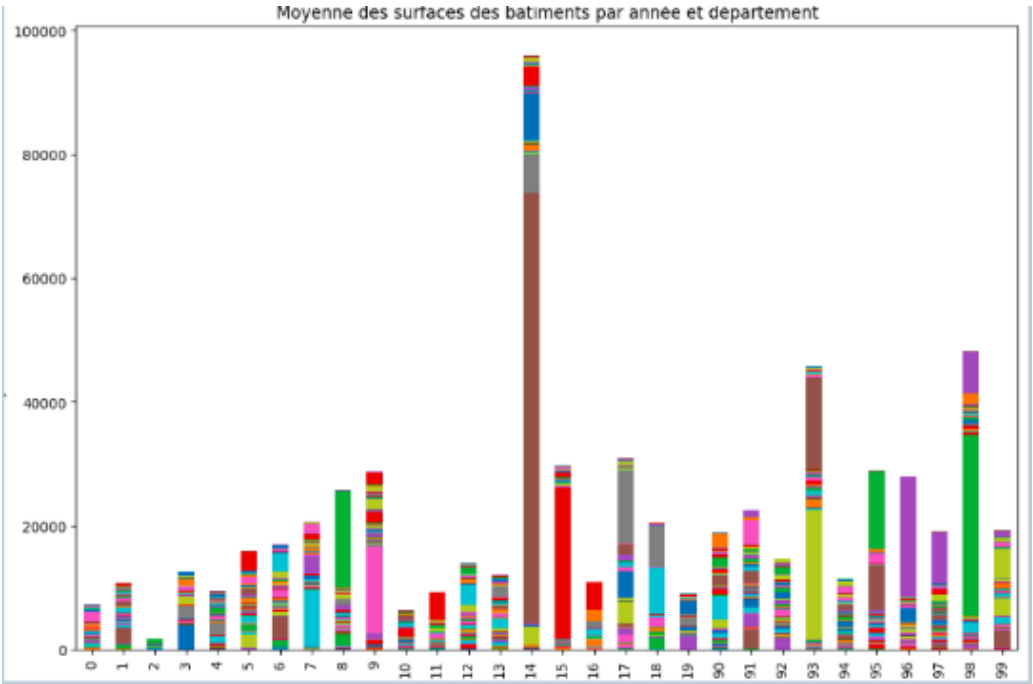


FIGURE 5

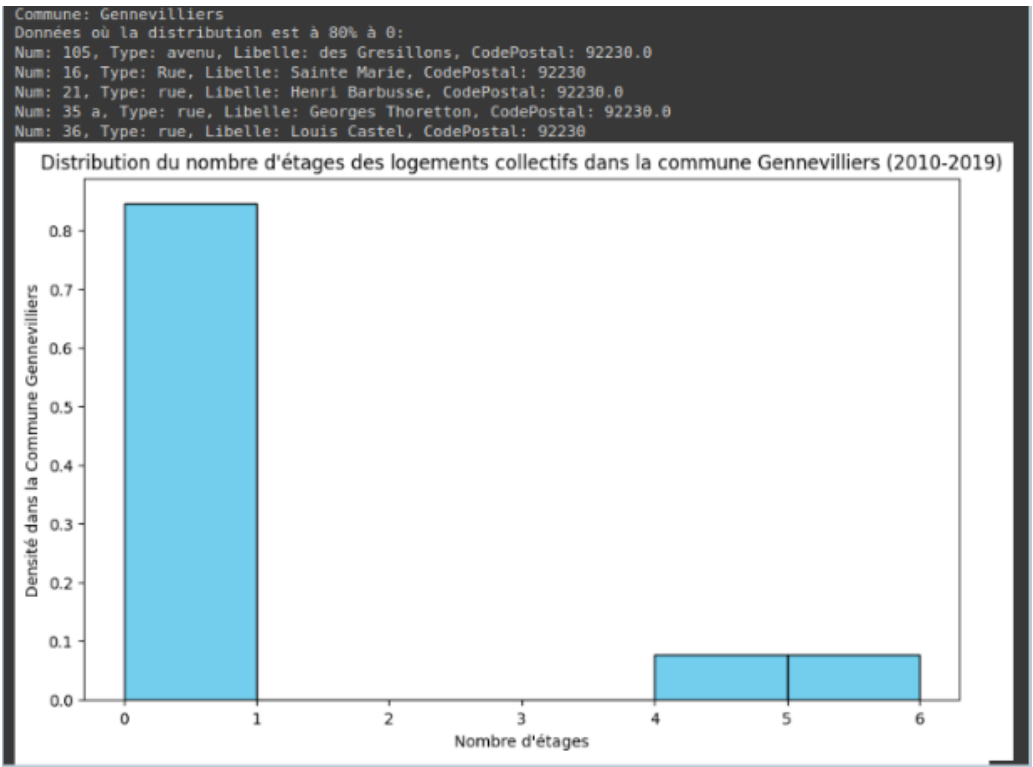


FIGURE 6

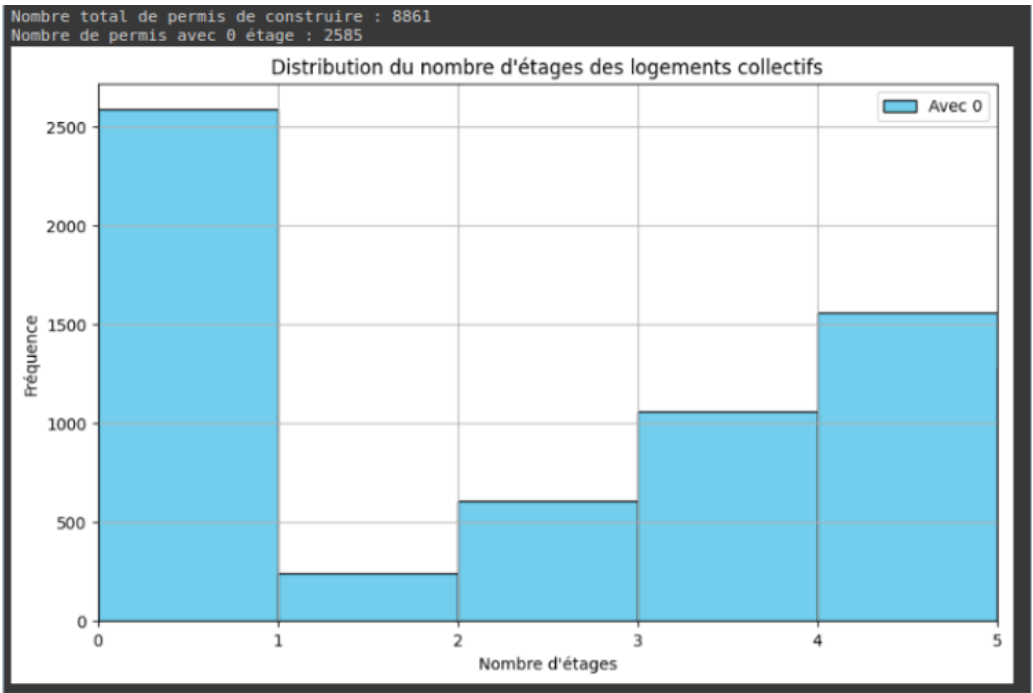


FIGURE 7