IUT DE PARIS

Axel CARTIER
STID 2018-2019

ANALYSE DE DONNÉES SPATIALES APPLIQUÉE AU PARC DE LOGEMENTS PARISIEN

Encadré par **Julien RANDON-FURLING**
Paris 1 : Panthéon-Sorbonne

Tutoré par **Olivier BOUAZIZ**
IUT Paris Descartes

Remerciements

Je tiens à remercier chaleureusement toute l'équipe du SAMM pour leur accueil ainsi que tous leurs collaborateurs. Encore mille mercis à Julien Randon-Furling de m'avoir suivi au cours de ces 10 semaines et à Jean-Marc Bardet de m'avoir accueilli au sein de son laboratoire.

Merci également aux personnes qui m'ont aidé à préparer et à relire ce dossier.

Sommaire

1. Introduction

Le dernier semestre du DUT [1] STID [2] en formation initiale est réservé aux stages. Ce stage est indispensable car il propose une approche concrète qui permet d'appliquer les connaissances acquises lors de notre formation.

Période de recherche J'ai d'abord postulé dans des entreprises du secteur de la banque-assurance car je souhaitais découvrir ce milieu. Plus tard, je me suis rendu compte que je portais un grand intérêt à l'enseignement, la diffusion et l'enrichissement des connaissances et c'est pourquoi j'ai préféré poursuivre mes recherches vers des établissements universitaires et des laboratoires de recherches.

Après plusieurs tentatives infructueuses, j'ai eu la chance de rencontrer, par le biais d'une connaissance commune, un enseignant chercheur qui travaille au laboratoire SAMM [3] à l'université Panthéon Sorbonne.

La mission C'est lors de notre première prise de contact que Monsieur Randon-Furling m'a décrit la mission qu'il allait me confier : programmer une méthode d'analyse spatiale développée au SAMM dans un langage informatique accessible.

La mission se précisa lors de mon arrivée au bureau. La première étape était de comprendre la méthode et de choisir quel langage de programmation j'allais utiliser. Ensuite, j'ai effectué quelques recherches puis j'ai programmé l'algorithme sur des données fictives avant de le tester sur la proportion de logements sociaux à Paris.

Ainsi, nous allons essayer de mieux comprendre le tissu social [4] parisien en observant la proportion de logements sociaux avec cette nouvelle méthode d'analyse spatiale.

1. Diplôme Universitaire et Technologique
2. Statistiques et Informatique Décisionnelle
3. Statistiques Analyse Modélisation Multidisciplinaire
4. Ensemble des interactions entre les individus

2. Présentation de l'environnement de travail

2.1 L'université Paris 1 Panthéon Sorbonne : un pôle universitaire historique et mondial

2.1.1 Bref historique [1]

C'est après les événements historiques de mai 68 que l'université de Paris fut divisée en 13 universités nouvelles. Sous l'impulsion des professeurs Hélène Ahrweiler (sciences humaines), François Luchaire (droit public), Henri Bartoli (économie) l'université Paris 1 Panthéon-Sorbonne voit le jour en 1971 avec le regroupement d'une partie de la faculté de droit et de sciences économiques (Panthéon) et d'une partie de la faculté des lettres et sciences humaines (Sorbonne). Depuis, les effectifs en perpétuelle augmentation de Paris 1 et la diversification des disciplines font que l'université a besoin d'agrandir ses locaux. Elle acquiert notamment le centre Saint-Charles et Tolbiac en 1973 et elle participe encore aujourd'hui à des projets de construction comme le campus de Condorcet ou celui de Port Royal.

2.1.2 Panthéon Sorbonne en chiffres [2]

Diffusion du savoir L'université reçoit plus de 42 000 étudiants et auditeurs chaque année dont 8 000 sont des étudiants étrangers en mobilité. Elle regroupe plus de 740 chercheurs et enseignants chercheurs dont les enseignements portent sur 3 principaux domaines : la gestion et l'économie, l'art et les sciences humaines et les sciences juridiques et politiques.

Recherche Paris 1 collabore avec 36 équipes de recherches, 10 écoles doctorales et 3 601 doctorants. 360 thèses et 40 HDR [1] ont été soutenues en 2015.

Chiffres divers L'institution bénéficie de plus de 25 implantations en Ile de France et elle est dirigée par un président, 8 vice-présidents, 2 vice-présidents étudiants et 30 personnes qui siègent au conseil d'administration.

1. Habilitation à Diriger des Recherches

2.1.3 Interactions avec les autres acteurs

Ambitions mondiales L'établissement a pour objectif de former de nombreux étudiants dans des domaines diversifiés que ce soit en présentiel ou à distance. C'est effectivement un pôle de savoir qui ambitionne de conquérir la scène mondiale comme l'atteste sa devise « Hic et ubique terrarum » qui signifie « Ici et partout sur la Terre ».

Partenaires L'université a des partenaires nationaux comme l'association française de sociologie et l'aéroclub de France mais elle appartient aussi à de vastes réseaux universitaires internationaux comme l'European university fondation, UNIMED [2] ou encore United Nations Academic Impact. Elle est partenaire de 294 établissements.

Concurrents Elle est en concurrence avec les autres universités françaises mais son rayonnement international en fait une concurrente des autres grandes universités mondiales. En effet, la Sorbonne est 91ème-100ème ex-aequo au World University Rankings du Times Higher Education de 2016, devancée par l'Ecole Normale Supérieure et l'université Pierre et Marie Curie.

2.2 Le laboratoire Statistiques Analyse et Modélisation Multidisciplinaire [3]

2.2.1 Présentation du laboratoire

Le SAMM est issu de la fusion en 2010 entre une équipe de recherche en mathématiques avec une autre spécialisée en économie et mathématiques financières. Il se situe au 20ème étage de l'institut Pierre Mendès France, rue Tolbiac à Paris.

Effectifs Le service comprend une cinquantaine de personnes dont 8 professeurs, 13 maîtres conférenciers, un PRAG [3], une chargée de gestion et 12 doctorants ou jeunes docteurs, en plus de la vingtaine de chercheurs associés.

Champs de recherches Les domaines de recherches couvrent de nombreux champs des mathématiques appliquées. Plus précisément, on trouve deux axes de recherches principaux au SAMM : « Statistique, Apprentissage Statistique et Réseaux » et « Dynamique et contrôle optimal ».

2. Universités de Méditerranée
3. Professeur agrégé

2.2.2 Partenariats variés et multiples

A l'échelle nationale Le laboratoire fait partie de la Fondation Sciences Mathématiques de Paris, du réseau Ile de France des Sciences de la cognition ainsi que de l'Institut des Systèmes Complexes d'Ile de France. Il est aussi impliqué dans de nombreux projets pluridisciplinaires notamment avec des archéologues, des géographes, des informaticiens et des économistes (Géographie-Cités, Pôle Informatique de Recherche et d'Enseignement en Histoire, Laboratoire Géographie-Physique...). Il a également développé de multiples liens avec des entreprises comme Viadeo, Orange Lab et bien d'autres encore.

A l'échelle internationale Le SAMM est aussi au cœur de projets internationaux : il collabore étroitement avec l'université de la Havane, le laboratoire de modélisation stochastique [4] et traitement de données d'Alger (Université Houari Boumedienne) puis également avec le réseau STAFAV [5].

2.3 Environnement humain

2.3.1 Gestion et organisation du laboratoire

Jean-Marc Bardet Jean-Marc est le directeur du laboratoire. Ingénieur de formation, il soutient une thèse en mathématiques en 1997. Après être devenue maître conférencier à Toulouse, il passe son HDR en 2002 et est recruté au SAMM un an après en tant que professeur des universités. Il devient directeur en 2012 et ses attributions sont multiples. Il se charge des questions logistiques et c'est lui qui fait le lien entre les différentes composantes de l'organisation et le laboratoire (Annexe I).

Diem Do Diem est la chargée de gestion du laboratoire. Elle s'occupe de la logistique lors du déplacement des membres du laboratoire mais aussi de l'accueil des intervenants invités. Elle se charge du budget de fonctionnement du laboratoire et elle gère également tous les achats (fournitures, consommables ...). Une grosse partie de sa charge de travail reste administrative : c'est aussi elle qui gère les conventions de recherche avec le service dédié de l'université et répond au téléphone pour le laboratoire.

2.3.2 Le bureau C.20.10

Julien Randon-Furling Julien est mon tuteur au laboratoire. Après une classe préparatoire en mathématiques-physique à Louis le Grand, il décroche une bourse à l'uni-

4. Aléatoire
5. Statistique pour l'Afrique Francophone et Applications au Vivant

versité de Cambrige où le laboratoire de mathématiques est profondément en lien avec celui de physique théorique. C'est lors de son cursus à Cambridge qu'il effectue deux stages sur le campus d'Orsay où il soutiendra sa thèse par la suite : « Statistiques d'extrêmes du mouvement brownien et applications »[4]. Il fait ensuite un an de post-doc à Sarrebruck où il travaille sur la physique statistique des systèmes complexes.

Puis Il est directement recruté au SAMM de Paris 1 en tant que maître de conférences et reprend des problématiques de géométrie stochastique qui ont été défrichées lors de sa thèse. Aujourd'hui, Julien enseigne les probabilités et les statistiques et travaille sur deux principaux thèmes : la modélisation stochastique et l'analyse de la ségrégation.

Aymen Hammami Aymen est un jeune docteur ayant fait sa thèse sur l'analyse harmonique [6] à la faculté El Manar de Tunis en partenariat avec la faculté d'Orsay. Il dispense des cours d'algèbre et d'analyse et ses recherches portent sur cette première discipline.

Antoine Lucquiaud Après une classe préparatoire en physique chimie, Antoine intègre l'ENS [7] puis passe l'agrégation de physique à Jussieu. Antoine prépare sa thèse au SAMM en rapport avec les modèles de Shelling [8]. Il dispense des travaux dirigés de probabilités pour les premières années de licence.

Alex Mourer Alex a effectué son cursus en MASS [9] à Paris 1, il prépare une thèse CIFRE [10] avec Safran et ses travaux portent sur la détection d'anomalies en bout de chaîne de production.

Cécile de Bézenac Cécile est diplômée d'un master en aménagement et développement durable qu'elle a effectué à l'école Centrale de Paris. Elle part ensuite à la faculté d'Aix-Marseille où elle intègre un master MIASH [11] et c'est dans le cadre de son master qu'elle collabore avec le SAMM.

6. Branche des mathématiques s'appuyant sur les notions de série de Fourrier et de transformée de Fourrier
7. Ecole Normale Supérieure
8. Modèles de ségrégation fondés dans les années 1970
9. Mathématiques Appliquées aux Sciences Sociales
10. Convention Industrielle de Formation par la Recherche
11. Mathématiques et Informatique Appliquées aux Sciences Sociales et Humaines

3. Cartographie de l'architecture des données

3.1 Source des données

3.1.1 Recensement de l'INSEE [5]

Les données que j'ai utilisées proviennent du recensement de 2013. Le recensement est une enquête effectuée par l'INSEE [1] à grande échelle sous la responsabilité de l'Etat et en coopération avec les EPCI [2]. Le recensement repose désormais sur une collecte annuelle d'informations concernant successivement tous les territoires communaux, sur une période de 5 ans. Sur cette période, environ 40 % des ménages [3] sont sondés. Ces informations sont ramenées au premier janvier de l'année médiane [4] pour toutes les communes afin d'assurer une meilleure robustesse des données.

Avant le début de la récolte, les communes recrutent des agents recenseurs que l'INSEE forme et cet organisme constitue le répertoire d'adresses à recenser. Après un repérage, les agents recenseurs diffusent aux ménages sondés des questionnaires papier ou des identifiants en ligne pour remplir le questionnaire. C'est la commune qui vérifie la bonne prise en compte de tous les ménages recensés et elle transmet les résultats à la direction régionale de l'INSEE qui effectue déjà de nombreux contrôles à cette étape.

Saisie des données Les résultats de l'enquête sont divisés en UT [5] (160 000 questionnaires environ) qui sont regroupés en LS [6](650 000 questionnaires environ). Ces LS sont envoyés à un prestataire d'acquisition des données qui scanne les bulletins puis constitue un fichier de données et des bases images. Un autre prestataire est chargé en parallèle de mesurer la qualité des bases images à l'aide d'un échantillon de contrôle. Après un examen approfondi effectué par par le pôle national pour les observations posant problème, les questionnaires sont renvoyés aux directions régionales de l'INSEE.

1. Institut national de la statistique et des études économiques
2. Établissement public de coopération intercommunale
3. Un foyer au sens du recensement
4. Celle qui se situe au milieu
5. Unités de Traitement
6. Lots de saisies

Codification Après cette étape, les fichiers de saisies sont repris et certaines variables sont codées automatiquement selon une nomenclature officielle formelle. Plus précisément, l'INSEE utilise la mise en concordance automatique pour le chiffrement de l'activité de l'établissement employeur. Il utilise d'autre part SICORE pour le chiffrement du pays, de la nationalité, de la commune, de la profession et de l'activité antérieure. Cependant, certains lots doivent être repris manuellement [7] et l'INSEE effectue également un contrôle qualité à cette étape. Ces traitements sont pour la plupart réalisés en batch processing [8].

Redressement Ces étapes permettent de corriger les données brutes lorsqu'elles sont incohérentes et de redresser les non réponses totales ou partielles.

On compte en tout trois redressements : le redressement des FLNE [9], le redressement de l'exploitation principale et le redressement de l'exploitation complémentaire. L'INSEE utilise un Hot Deck [10] pour le redressement d'une variable non renseignée alors qu'il utilise une imputation déterministe pour redresser les incohérences. Ces traitements sont aussi effectués en batch processing.

Re-codification et validation A l'issue du redressement automatique de toutes les variables, il faut encore les mettre en forme pour une meilleure compréhension des résultats. Certaines variables permettent de synthétiser d'autres existantes soit en regroupant les informations soit en se limitant à deux ou trois modalités. Certaines variables expriment notamment une durée et sont calculées à partir d'une date. D'autres variables synthétisent plusieurs autres variables existantes.

Mise à disposition des données Les résultats statistiques du recensement "n" sont diffusés au cours du premier semestre "n+3" alors que les populations légales du millésime "n" sont diffusées fin décembre de l'année "n+2". Toutes les données sont rendues anonymes comme l'exige la RGPD [11] puis rendues disponibles directement sur le site de l'INSEE pour enrichir l'open data [12].

7. Environ 1% en métropole
8. Traitement par lots, enchaînement automatique d'une suite de commandes sans l'intervention d'un opérateur. Une fois le processus terminé, l'ordinateur l'applique au lot suivant.
9. Fiche de Logement Non Enquêté
10. Imputation par le plus proche voisin
11. Règlementation Générale sur la Protection des Données
12. Données ouvertes, données numériques dont l'accès et l'usage sont laissés libres aux usagers

3.1.2 Découpage infracommunal français : les IRIS

Définition [6] En 1999, afin de préparer la diffusion du recensement, l'INSEE développe un découpage plus fin du territoire de taille homogène que l'on nommait IRIS [13]2000 [14].

Aujourd'hui, l'IRIS est devenu la brique de base en matière de diffusion de données infracommunales : toutes les communes d'au moins 10 000 habitants et la plupart des communes de 5 000 à moins de 10 000 habitants sont découpées en IRIS. La France compte environ 16 100 IRIS dont 650 dans les DROM [15].

Afin de couvrir tout le territoire, on assimile à un IRIS toutes les communes non découpées en IRIS. On dénombre 3 types d'IRIS : Habitat, Activité et Divers. Certains IRIS sont regroupés par trois, ce que l'on nomme un TRIRIS.

Intégration avec les données de l'INSEE Les données publiées à l'année "n" se basent sur la découpe géographique faite au 1er janvier de l'année "n+2". Les données de l'enquête logement de 2013 sont donc basées sur la découpe du 1er janvier 2015. Les bases de données contenant les informations sur les IRIS sont disponibles sur le site de l'IGN [16].

3.2 Outils utilisés et environnement technique

3.2.1 Choix du langage de programmation

Madalina Olteanu a déjà codé cet algorithme avec le langage de programmation R [17] lors de la publication de l'article[7], c'est pourquoi j'ai choisi Python [18]. De plus, Python tend à se démocratiser autant dans le milieu de la recherche que celui des entreprises.

3.2.2 Environnement informatique

Spécifications matérielles et système d'exploitation J'ai utilisé mon ordinateur personnel pour cette mission. Il s'agit d'un Macbook Pro de 2013 disposant de 16Go de DDR3 SDRAM [19], d'un processeur Intel Core i7 cadencé à 2,3GHz et de deux processeurs graphiques : une carte NVIDIA GeForce GT 750M 2 Go appuyée par une Iris Pro

13. Ilot Regroupé pour l'Information Statistique
14. Le chiffre fait référence à la taille visée de 2000 habitants par maille élémentaire
15. Départements et Régions d'Outre Mer
16. Institut national de l'information géographique et forestière
17. Langage de programmation créé pour faciliter les analyses statistiques
18. Langage de programmation et logiciel libre multiplateforme favorisant la programmation impérative structurée, fonctionnelle et orientée objet
19. Double Data Rate 3rd generation Synchronous Dynamic Random Access Memory

1536 Mo de la marque Intel. Cet ordinateur fonctionne avec la version 10.14.5 de MacOS [20] Mojave.

Logiciels Souhaitant programmer cet algorithme en Python, je me suis muni de la dernière version de l'environnement Anaconda [21] et j'ai utilisé le logiciel Spyder [22] pour construire mon programme. Pour l'import, la préparation et la visualisation des données du recensement, je me suis servi du logiciel Rstudio [23] (Annexe II).

D'autre part, je me suis servi de MindView [24], de la suite Microsoft Office, d'Adobe Photoshop ainsi que de TeXShop installé via l'environnement MacTeX pour construire ce rapport.

Librairies utilisées Sur l'environnement R, je ne me suis servi que du package dplyr qui offre une syntaxe agréable pour la manipulation des données. Outre les packages nécessaires à la manipulation des données sous Python comme Numpy, Pandas et Matplotlib, j'ai également utilisé des packages plus spécifiques comme Geopandas[8] pour représenter les données géolocalisées et Scikit-Learn [25] pour certains calculs plus poussés.

3.3 Organisation des bases de données (Annexe III)

3.3.1 Données brutes

Logements ordinaires 2013 [9] Cette base de données regroupe ainsi les informations concernant les logements ordinaires selon la définition du recensement : c'est un local utilisé pour l'habitation, clos et indépendant. Ne sont pas considérés comme logements ordinaires les habitations mobiles et les locaux habités par des personnes résidant au sein d'une communauté. Ce jeu de données contient 70 variables et 24 256 179 observations. Cependant, cette base de données est disponible découpée en 5 zones et j'ai donc choisi celle correspondant à l'Ile de France qui contient 2 694 072 entrées. Ici, les variables qui nous intéressent sont le poids du logement dans l'IRIS (IPONDL), le code

20. Macintosh Operating System
21. Distribution libre et open source des langages de programmation Python et R appliquée au développement d'applications dédiées à la science des données et à l'apprentissage automatique visant à simplifier la gestion des paquets et de déploiement. Il contient notamment les logiciels Spyder et Rstudio.
22. Environnement de développement sur Python contenant l'accès aux nombreuses bibliothèques scientifiques
23. Environnement de développement R facilitant l'intégration des paquetages
24. Logiciel permettant de créer des cartes mentales
25. Paquetage implémentant de nombreux outils pour l'analyse de données comme des algorithmes de classification, de régression, de clustering ...

de l'IRIS et la variable HLML qui indique si un logement appartient à un organisme HLM [26] ou non.

Découpage géographique [10] Le shapefile [27] contient les données servant à cartographier la France. On trouve ainsi 50 152 observations dont chacune représente un IRIS pour 7 variables : le code commune de l'INSEE, le nom de la commune, le nom, le type et la géométrie de l'IRIS ainsi que deux codes associés.

3.3.2 Tables intermédiaires

Paris En premier lieu, j'ai filtré les observations du shapefile pour garder uniquement les IRIS parisiens à l'aide de la recherche du terme "Paris" dans le nom des communes. Cependant, après vérification, il a fallu supprimer les communes de Parisot et de Paris-l'Hôpital. J'ai délibérément retiré les deux bois parisiens et le jardin du Luxembourg car ce ne sont pas des IRIS dédiés à l'habitation. Pour finir, j'ai isolé les coordonnées en abscisse et en ordonnée de l'isobarycentre [28]de chaque IRIS à l'aide des attributs "centroïd.x" et "centroïd.y".

Log Paris F D'autre part, j'ai restreint les données concernant les logements uniquement aux logements déclarés comme résidence principale. Ensuite, j'ai sommé les poids des logements appartenant à un organisme HLM en fonction de l'identifiant de l'IRIS pour obtenir le nombre de logements sociaux par IRIS. En parallèle, j'ai fait de même en sommant les poids de toutes les résidences principales en fonction de l'identifiant de l'IRIS pour obtenir le nombre de logements au total. Pour finir, il m'a suffi de fusionner ces deux tables en fonction de l'identifiant de l'IRIS, de renommer les variables et de remplacer les valeurs manquantes par 0.

3.3.3 Table finale

Après avoir fusionné les données du recensement avec les informations du shapefile et conservé les variables nécessaires, j'ai supprimé les IRIS qui n'ont pas de logements puis j'ai défini ensuite le taux de logements sociaux pour chaque IRIS comme étant le nombre de logements sociaux divisé par le nombre de logements au total dans un IRIS.

Puis, j'ai créé 4 variables vides qui contiendront par la suite une liste pour chacune des observations. Une contiendra l'indice des plus proches voisins ordonné dans l'ordre

26. Habitation à Loyer Modéré
27. Fichier de forme, format de fichier standard pour les systèmes d'informations géographiques initialement développé par ESRI
28. En mathématiques, cette notion généralise la notion de centre de gravité pour un triangle

croissant de la distance, deux contiendront respectivement le nombre de logements au total et celui de logements sociaux pour chaque niveau d'agrégation et la dernière contiendra le taux de logements sociaux pour chaque niveau d'agrégation.

Pour finir, j'ai ajouté une variable vide que j'ai appelée "check" qui contiendra le niveau d'agrégation à partir duquel on considère que le taux de logements sociaux converge vers la moyenne parisienne.

4. Analyse de la ségrégation spatiale : part des logements sociaux à Paris

4.1 Enjeux et objectifs

Le SAMM étant en étroit lien avec les chercheurs en sciences humaines et sociales de Paris 1, Julien, aidé de Madalina Olteanu et d'Antoine Lucquiaud, ont développé une nouvelle méthode d'analyse basée sur les plus proches voisins pour détecter les dissemblances dans les structures spatiales. Dans leur article[7], ils testent leur nouvelle méthode pour mettre en lumière la ségrégation [1] spatiale à Paris, plus précisément avec la proportion de logements HLM. L'objectif de ma mission est de reproduire les résultats de l'article en comprenant et en programmant cette méthode avec un autre langage de programmation que celui qui a été utilisé. Pour atteindre cet objectif, je me suis muni d'un calendrier prévisionnel ainsi que d'un cahier des charges (Annexe IV).

4.2 Méthodologie employée

4.2.1 Description détaillée de la méthode

Pour calculer les disparités spatiales, nous allons partir d'un IRIS pour agréger un par un le nombre de logements et celui de logements sociaux en fonction de la proximité avec les autres IRIS : on commence par les plus proches. Après avoir fixé préalablement un seuil, nous regarderons le taux de logements sociaux pour tous les niveaux d'agrégation et pour chaque IRIS. Nous relèverons le niveau d'agrégation à partir duquel le taux de logements sociaux entre à l'intérieur des bornes fixées par le seuil et on considèrera qu'il converge vers la moyenne globale à partir de ce niveau d'agrégation.

Calcul des bornes de convergence Nous nous basons sur le taux de logements sociaux dans Paris : nous prenons la somme du nombre de logements sociaux à Paris divisé par la somme du nombre de logements au total. Ensuite nous choisissons un seuil, ici 5%. Les bornes de l'intervalle seront données par le taux global diminué de 5% d'une part et augmenté de 5% d'autre part.

1. Désigne tout phénomène évolutif ou tout état de séparation de groupes ethniques ou sociaux

Recherche du plus proche voisin [11] Afin d'obtenir une liste contenant l'indice des plus proches voisins dans l'ordre croissant, j'ai utilisé l'algorithme KDTree [2] de Scikit-Learn. En effet, ce dernier prend en entrée un jeu de données et calcule les "n" plus proches voisins pour chaque observation. En cas d'égalité de distance avec deux points, l'algorithme la casse arbitrairement en fonction de l'index de l'observation dans le data frame. C'est pourquoi j'ai décidé de randomiser [3] l'index de mon jeu de données en amont pour pallier ce problème. Par défaut, l'algorithme travaille avec la distance euclidienne [4] (Figure 4.1) mais il est possible de modifier ce paramètre. On peut ensuite interroger cet objet pour obtenir l'indice des "n" plus proches voisins de telle observation dans l'ordre croissant de la distance.

Méthode d'Agrégation Il a fallu créer trois suites pour chaque IRIS : une contenant le nombre de logements au total pour tous les niveaux d'agrégation, une avec le nombre de logements sociaux pour tous les niveaux d'agrégation et une suite se basant sur les deux précédentes, permettant de calculer le taux de logements sociaux toujours pour chaque niveau d'agrégation.

4.2.2 Cheminement au cours du développement (Annexe V)

Compréhension Ma première tâche a été de comprendre cette méthode. Julien m'a ainsi transmis son article imprimé avec l'annexe concernant la méthode d'agrégation. Après avoir lu et relu cet article de manière approfondie, j'ai pu discuter avec Julien afin d'éclaircir les dernières zones d'ombres.

Recherches La deuxième phase était celle des recherches : je cherchais les algorithmes existants en Python desquels je pouvais m'inspirer pour rechercher le plus proche voisin. Je me suis également intéressé aux algorithmes de remplissage, de path-finding [5] et aux librairies Python spécialisées dans le traitement des données géographiques comme Geopandas et libpysal.

Tests Lors de la phase suivante, j'ai manipulé des données générées aléatoirement sur une grille. Effectivement, je me suis entraîné sur une grille 10x10 où j'avais attribué un nombre aléatoire de logements sociaux et un nombre aléatoire de logements obligatoi-

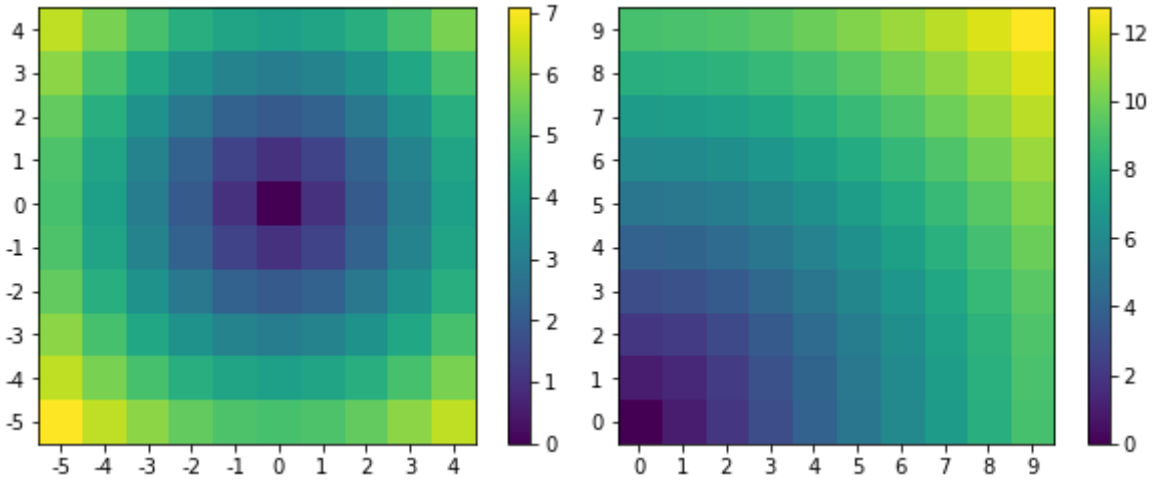
2. Arbre kd, structure de données permettant notamment de faire efficacement des recherches des plus proches voisins

3. Permutation aléatoire d'une séquence

4. Distance donnée par $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ pour deux points x et y dans un espace métrique à n dimension(s)

5. Algorithme de recherche du plus court chemin comme l'algorithme de Dijkstra

FIGURE 4.1 – Distance euclidienne à (0,0)



rement plus grand que le premier pour chaque case de la grille. Cette étape m'a permis de comprendre comment j'allais procéder par la suite.

Développement En premier lieu, j'ai développé une classe IRIS à laquelle j'ai greffé des méthodes pour calculer l'index des plus proches voisins. Cependant, cette approche devenait compliquée au moment où il fallait la généraliser pour n'importe quel jeu de données géographiques. C'est pourquoi j'ai choisi de complètement changer de cap pour utiliser l'objet KDTree de Scikit-Learn. Il me suffisait ainsi d'utiliser une double boucle pour parcourir tous les éléments du jeu de données et les indices des plus proches voisins. J'ai construit ainsi le squelette de mon programme en différentes étapes : préparation des données, calcul du seuil et randomisation des indices, recherche de l'index des plus proches voisins, calcul des niveaux d'agrégation et enfin rendus graphiques.

Débogage Toutefois, lorsque le programme fut terminé, les résultats obtenus ne concordaient pas avec les résultats de l'article. Ont suivi quelques jours de tests avec une grille pour comprendre d'où venait le problème. Au final, c'est une condition abusive qui écrasait le niveau d'agrégation où l'on considère que la suite obtenue converge.

4.3 Résultats

4.3.1 Résultats de l'analyse

Taux de logements sociaux par IRIS Nous pouvons ainsi représenter sur une carte le taux de logements sociaux pour chaque point de départ. On voit ainsi que certains

IRIS sont composés quasi-uniquement de logements sociaux alors que d'autres n'en comptent pas un seul (Figure 4.3).

Rapidité de convergence vers la moyenne globale Pour observer cet aspect, j'ai utilisé une carte dont la couleur d'un IRIS est en relation avec le niveau d'agrégation nécessaire pour que le taux converge vers la moyenne générale (Figure 4.4).

Ensuite, j'ai pu représenter pour chaque IRIS une courbe avec en abscisse le niveau d'agrégation et en ordonnée le taux de logements sociaux. Une barre verticale bleue représente le moment où une des suites converge vers la moyenne et les barres horizontales noires représentent les bornes du seuil (Figure 4.2).

Analyse détaillée Pour avoir plus de précisions sur la convergence des IRIS, je me suis intéressé aux indicateurs statistiques de la variable dédiée. En moyenne, il faut 446 agrégations avant que la convergence ait lieu et la distribution est plutôt dispersée (écart type de 382,405589). La moyenne est relativement proche de la médiane qui, elle, vaut 430. Il faut entre 0 et 922 agrégations avant que le taux converge vers la moyenne globale. 25% des IRIS nécessitent 20 agrégations au plus avant la convergence et 25% nécessitent au moins 895 agrégations avant de rejoindre le taux parisien. L'IRIS

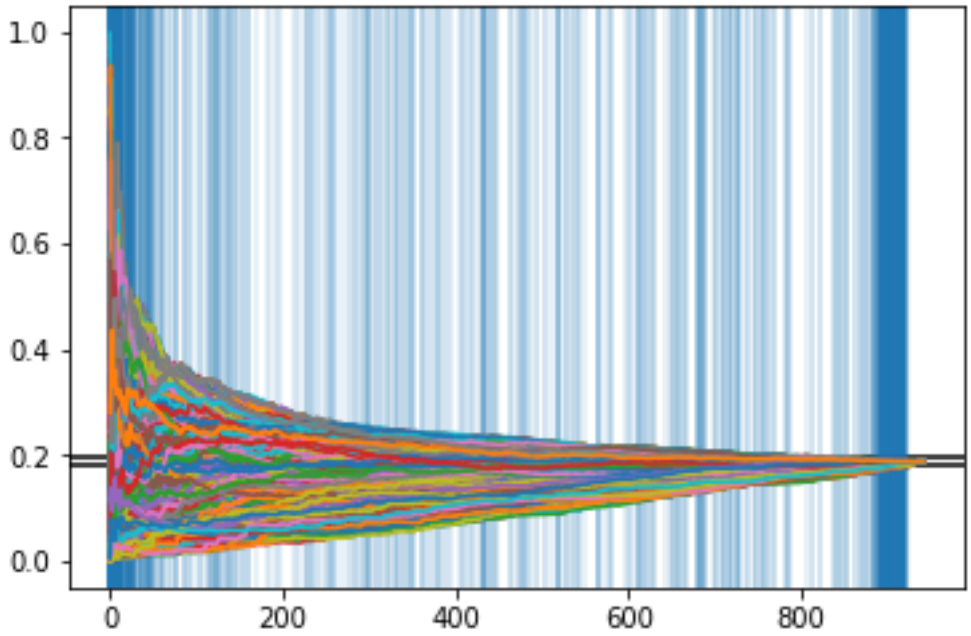


FIGURE 4.2 – Trajectoires du taux de logements sociaux en fonction du niveau d'agrégation

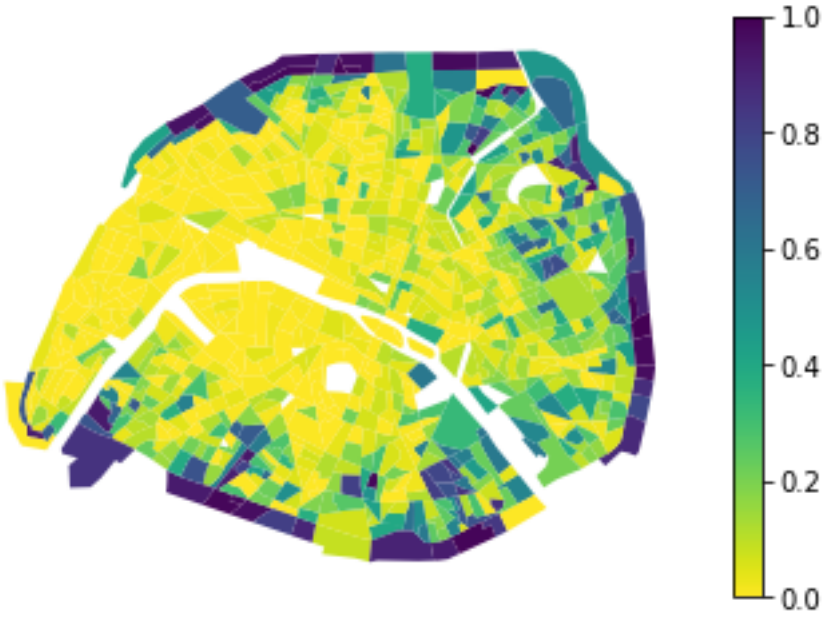


FIGURE 4.3 – Taux de logements sociaux par IRIS

FIGURE 4.4 – Rapidité de convergence vers la moyenne globale

qui converge le plus lentement est situé dans le 16ème arrondissement (Auteuil 22) et plusieurs IRIS ont un taux similaire au taux global parisien notamment dans les 5ème, 9ème, 11ème, 12ème, 13ème, 14ème, 15ème, 18ème et 20ème arrondissements.

On remarque ainsi que le sud-ouest parisien semble compter moins de logements sociaux que dans le reste de la capitale. En effet, c'est une zone extrêmement attractive pour les ménages aisés et la forte demande fait augmenter le prix des logements du quartier, ce qui favorise l'exclusion des ménages les plus pauvres. Pour Anne Clerval, le parc parisien de logements est profondément modifié par le processus de gentrification [6] qui contribue à expliquer cette ségrégation sociale[12].

4.3.2 Adéquation avec la formation

Ce stage complète parfaitement la formation STID. Il m'a permis de mettre en application la plupart des compétences apprises lors du DUT.

Développement informatique La première partie du stage consistait à programmer la méthode d'analyse développée par le laboratoire. Avant la programmation, il a donc fallu préparer le squelette du programme : j'ai réfléchi aux entrées et sorties du programme et je me suis fixé des objectifs à court terme pour m'auto-motiver.

Recherche et traitement des données La seconde partie du stage était dédiée à faire fonctionner le programme sur des données réelles. Je me suis donc imprégné du fonctionnement du site de l'INSEE pour pouvoir y naviguer convenablement et récupérer les informations nécessaires, que ce soit des bases de données, des dictionnaires de variables ou des précisions sur la méthodologie.

Une fois les données brutes récoltées, il fallait les manipuler pour obtenir les informations souhaitées. J'ai aussi fusionné celles-ci avec les données géographiques.

Enfin, j'ai analysé les résultats obtenus, notamment à l'aide d'indicateurs statistiques.

Restitution de l'information L'autre compétence travaillée au cours de ce diplôme est la restitution d'informations pour les décideurs. En plus du rapport servant de synthèse, j'ai souhaité trouver le meilleur rendu graphique permettant d'expliquer au mieux les résultats obtenus.

6. Embourgeoisement, phénomène urbain par lequel les populations aisées s'approprient un espace jusque là occupé majoritairement par des habitants moins favorisés, jusqu'à transformer l'espace en question

4.3.3 Limites de la méthode

Toutefois, cette méthode génère des freins qui limitent son utilisation. Certains sont dus à la façon dont ont été récoltées les données, d'autres surviennent lorsque le maillage du territoire n'est pas assez fin et la trajectoire parcourue lors de l'agrégation reflète mal l'exploration des environs d'une personne vivant à Paris.

Récolte des données Les données du recensement sont recueillies par un questionnaire et les réponses sont uniquement déclaratives. L'INSEE appelle alors à la prudence concernant la fiabilité des estimations liées au dénombrement des logements sociaux[13].

Maillage grossier Si les données sont déjà regroupées dans de gros agrégats, les unités de base vont statistiquement être plus proches de la moyenne que si le maillage territorial était fin. Ainsi, les différences de convergence vers la moyenne globale seront moins marquées.

Trajectoire d'exploration La façon dont l'algorithme agrège les IRIS n'est pas représentative de la manière dont un Parisien explorerait la ville. En effet, on aura tendance à beaucoup explorer les environs de notre domicile et ceux de notre lieu de travail comparé au reste de la ville. Nous prenons en compte une seule trajectoire par point de départ alors qu'il en existe des millions.

4.4 Réflexion et prise de recul sur la mission

4.4.1 Apports de la mission et compétences sollicitées

Ce stage m'a beaucoup apporté sur plusieurs plans. Il m'a d'abord permis d'améliorer mes compétences en traitement des données en plus d'apprendre de nouveaux savoir-faire. D'autre part, il m'a permis de me familiariser avec le monde de la recherche et de développer ma manière de communiquer et ma rédaction.

Compétences techniques Je me suis formé à l'utilisation de multiples librairies Python pour réaliser cette étude. En effet, j'ai découvert les packages servant à traiter des données spatiales et géométriques comme Shapely, libpysal et Geopandas par exemple. J'ai également découvert une extension standardisée de fichier fort connue par les géographes que l'on nomme shapefile.

De manière transversale, j'ai également pu développer mes compétences en gestion de projet informatique et en restitution de l'information. J'ai notamment appris le langage LaTeX au cours de ce stage. Lors de ce projet, j'ai travaillé en autonomie la plupart du temps, ce qui m'a permis de développer mes capacités d'auto-gestion.

Autres bénéfices L'autre apport majeur est la découverte d'un laboratoire de recherches. En effet, je me suis familiarisé avec certains standards académiques comme le fameux "papier" [7] et les différents statuts qui existent au sein de la profession.

De plus, le laboratoire organise tous les vendredis matin une conférence sur des sujets tous plus intéressants les uns que les autres. Par exemple, Andreas Kerren est venu nous présenter une conférence sur les différentes techniques de visualisation de données textuelles. Lê Nguyên Hoang est venu nous parler de la morale dans l'intelligence artificielle et Paul Raynaud de Fitte est venu de Rouen pour une conférence sur le processus de rafle perturbé par un signal rugueux.

Ceci m'a permis de faire de merveilleuses rencontres et d'avoir d'extraordinaires discussions avec des chercheurs.

4.4.2 Réflexion sur les difficultés rencontrées

La première difficulté à laquelle j'ai été confronté était de calculer la matrice des distances puis de récupérer l'indice des points les plus proches. Si j'arrivais à calculer la matrice des distances, je n'ai pas réussi à récupérer l'indice du point lié à une distance et c'est ce qui m'a poussé à modifier ma première approche et à utiliser des classes d'objets existantes.

7. Appellation pour un article scientifique

Ensuite, j'ai été confronté aux problèmes d'optimisation de mon programme. S'il tourne entre 5 et 15 min avec le taux de logements sociaux pour les 942 IRIS sélectionnés, l'algorithme sera beaucoup plus lent avec un gros jeu de données à cause de l'utilisation de fonctions dans des boucles imbriquées notamment.

4.4.3 Conclusion technique

En somme, ce projet très complet m'a permis de progresser dans beaucoup de domaines directement liés ou non aux statistiques. Même si j'ai changé à plusieurs reprises de méthode, j'ai réussi à fournir un programme adapté à la problématique qui m'était posée.

Pérennisation du projet Pour pérenniser ce programme, il faudrait le modifier légèrement pour pouvoir choisir le nom des variables à agréger et effectuer un traitement général préparatif pour le jeu de données en entrée.

Autres applications envisagées Dans la continuité de cette mission, nous envisageons de faire tourner ce programme avec les données des autres grandes villes françaises ou sur des données parisiennes fournies selon un maillage plus fin.

5. Conclusion

Lors de mon arrivée au bureau, j'étais très impressionné et je me demandais comment j'allais pouvoir réussir à programmer ce qui m'était demandé. Mais en prenant le temps de poser mes idées puis de les tester, je suis arrivé malgré les nombreux changements de bords, à produire des résultats concrets basés sur des données réelles. J'ai pu apprendre de nombreux savoir-faire et réviser mes acquis au cours de cette mission.

Un métier passionnant Dans l'imaginaire collectif, on trouve selon moi deux visions caricaturales du chercheur qui s'opposent : celle du chercheur solitaire devenu fou car on lui a volé ses travaux et une autre, idéalisée, où il prône le collectif et souhaite enrichir la connaissance commune. J'ai pu me rendre compte au cours de ce stage que la vérité se situe sûrement entre ces deux extrêmes. Une grande part du métier de chercheur est individuelle même s'il est régulièrement ponctué de nombreuses rencontres, collaborations, discussions, cours et séminaires en tous genres. C'est l'interaction entre les chercheurs et leurs différentes spécialités qui permet souvent de mieux comprendre le monde, de créer de nouveaux modèles et de repousser les frontières de la connaissance.

En m'intéressant au processus de la recherche scientifique, j'ai aussi pu me rendre compte de quelques aberrations comme le fait que les chercheurs doivent payer dans certains cas pour que leurs travaux soient publiés dans une revue scientifique reconnue qui, elle-même, doit être achetée par les universitaires. J'ai également déchanté lorsque j'ai parlé de rémunération avec les jeunes doctorants.

Toutefois, c'est un métier plus que passionnant, extrêmement enrichissant et nécessaire.

Perspectives d'avenir Je souhaite continuer mes études en licence de mathématiques appliquées pour ensuite repartir vers un parcours plus théorique. Le métier d'enseignant chercheur est pour moi un rêve et j'espère pouvoir continuer mes études jusqu'à le réaliser.

Bibliographie

[1] Paris 1 PANTHÉON-SORBONNE. *Historique de l'université*. https://www.pantheonsorbonne.fr/universite/presentation/historique. Consulté le 10-04-19. 2019.

[2] Paris 1 PANTHÉON-SORBONNE. *Chiffres clefs*. https://www.pantheonsorbonne.fr/fileadmin/Service-com/webkey/Universite/en_chiffre_2015_novembre_FR.pdf. Consulté le 25-04-19. 2015.

[3] Cédric COUVRAT. *Site web du SAMM*. http://samm.univ-paris1.fr. Consulté le 10-04-19. 2017.

[4] Julien RANDON-FURLING. "Extreme-Value Statistics of Brownian Motion, and Applications". Theses. Université Paris Sud - Paris XI, nov. 2009. URL : https://tel.archives-ouvertes.fr/tel-00524212.

[5] INSEE. *Le traitement des données du recensement de la population*. https://www.insee.fr/fr/information/2526415. Consulté le 17-05-19. 2019.

[6] INSEE. *Définition - IRIS*. https://www.insee.fr/fr/metadonnees/definition/c1523. Consulté le 20-05-19. 2016.

[7] Julien RANDON-FURLING, Madalina OLTEANU et Antoine LUCQUIAUD. "From Urban Segregation to Spatial Structure Detection". In : *Environment and Planning B : Urban Analytics and City Science* (déc. 2017).

[8] GeoPandas DEVELOPERS. *Documentation GeoPandas*. http://geopandas.org. Consulté le 02-05-19. 2019.

[9] INSEE. *Données logements ordinaires 2013*. https://www.insee.fr/fr/statistiques/2409491?sommaire=2409559. Consulté le 02-05-19. 2016.

[10] IGN. *Contours IRIS*. http://professionnels.ign.fr/contoursiris. Consulté le 02-05-19. 2017.

[11] scikit-learn DEVELOPERS. *Documentation ArbreKD*. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KDTree.html. Consulté le 13-04-19. 2019.

[12] Anne CLERVAL. *Paris sans le peuple. La gentrification de la capitale*. La Découverte, 2013. URL : https://hal-upec-upem.archives-ouvertes.fr/hal-00864885.

[13] INSEE. *Le dispositif statistique de l'Insee dans le domaine du logement - État des lieux et évaluation comparée des sources*. https://www.insee.fr/fr/statistiques/1380822. Consulté le 26-05-19. 2010.

Table des figures

Glossaire

A

Anaconda Distribution libre et open source des langages de programmation Python et R appliquée au développement d'applications dédiées à la science des données et à l'apprentissage automatique visant à simplifier la gestion des paquets et de déploiement. Il contient notamment les logiciels Spyder et Rstudio. 11

analyse harmonique Branche des mathématiques s'appuyant sur les notions de série de Fourrier et de transformée de Fourrier. 7

B

batch processing Traitement par lots, enchaînement automatique d'une suite de commandes sans l'intervention d'un opérateur. Une fois le processus terminé, l'ordinateur l'applique au lot suivant. 9

C

CIFRE Convention Industrielle de Formation par la Recherche. 7

D

DDR3 SDRAM Double Data Rate 3rd generation Synchronous Dynamic Random Access Memory. 10

distance euclidienne Distance donnée par $\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$ pour deux points x et y dans un espace métrique à n dimension(s). 15

dplyr Paquetage du tidyverse permettant de manipuler les données selon le modèle du tidy data. 11

DROM Départements et Régions d'Outre Mer. 10

DUT Diplôme Universitaire et Technologique. 3, 19

E

ENS Ecole Normale Supérieure. 7

EPCI Établissement public de coopération intercommunale. 8

F

FLNE Fiche de Logement Non Enquêté. 9

G

gentrification Embourgeoisement, phénomène urbain par lequel les populations aisées s'approprient un espace jusque là occupé majoritairement par des habitants moins favorisés, jusqu'à transformer l'espace en question. 19

Geopandas Paquetage implémentant la classe d'objets "geo data frame" permettant de gérer les jeux de données géographiques en Python. 11, 15, 21

H

HDR Habilitation à Diriger des Recherches. 4

HLM Habitation à Loyer Modéré. 12, 14

Hot Deck Imputation par le plus proche voisin. 9

I

IGN Institut national de l'information géographique et forestière. 10

INSEE Institut national de la statistique et des études économiques. 8–10, 12, 19, 20

IRIS Ilot Regroupé pour l'Information Statistique. 10–12, 14–17, 20, 22, 34, 35

isobarycentre En mathématiques, cette notion généralise la notion de centre de gravité pour un triangle. 12

K

KDTree Arbre kd, structure de données permettant notamment de faire efficacement des recherches des plus proches voisins. 15, 16

L

libpysal Librairie Python pour l'analyse spatiale. 15, 21

LS Lots de saisies. 8

M

MacOS Macintosh Operating System. 11

MacTeX Distribution TeX libre basée sur TeX Live, spécialement destinée à la plate-forme Mac OS X. 11

MASS Mathématiques Appliquées aux Sciences Sociales. 7

Matplotlib Paquetage facilitant la représentation graphique. 11

ménage Un foyer au sens du recensement. 8, 19

MIASH Mathématiques et Informatique Appliquées aux Sciences Sociales et Humaines. 7

N

Numpy Paquetage implémentant la classe d'objets "array" permettant de gérer les matrices en Python. 11

O

open data Données ouvertes, données dont l'accès et l'usage sont laissés libres aux usagers. 9

P

Pandas Paquetage implémentant la classe d'objets "data frame" permettant de gérer les bases de données en Python. 11

PRAG Professeur agrégé. 5

Python Langage de programmation et logiciel libre multiplateforme favorisant la programmation impérative structurée, fonctionnelle et orientée objet. 10, 11, 15, 21

R

R Langage de programmation libre créé pour faciliter les analyses statistiques. 10, 11

randomiser Permuter aléatoirement une séquence. 15

recensement Enquête exhaustive de la population menée par l'INSEE. 8, 9, 11, 20

RGPD Règlementation Générale sur la Protection des Données. 9

Rstudio Environnement de développement R facilitant l'intégration des paquetages. 11

S

SAMM Statistiques Analyse Modélisation Multidisciplinaire. 3, 5–7, 14

Scikit-Learn Paquetage implémentant de nombreux outils pour l'analyse de données comme des algorithmes de classification, de régression, de clustering ... 11, 15, 16

ségrégation Désigne tout phénomène évolutif ou tout état de séparation de groupes ethniques ou sociaux. 7, 14, 19

shapefile Fichier de forme, format de fichier standard pour les systèmes d'informations géographiques initialement développé par ESRI. 12, 21

Spyder Environnement de développement sur Python contenant l'accès aux nombreuses bibliothèques scientifiques. 11

STAFAV Statistique pour l'Afrique Francophone et Applications au Vivant. 6

STID Statistiques et Informatique Décisionnelle. 3, 19

stochastique Aléatoire. 6, 7

T

TeXShop Editeur TeX pour MacOS. 11

tissu social Ensemble des interactions entre les individus. 3

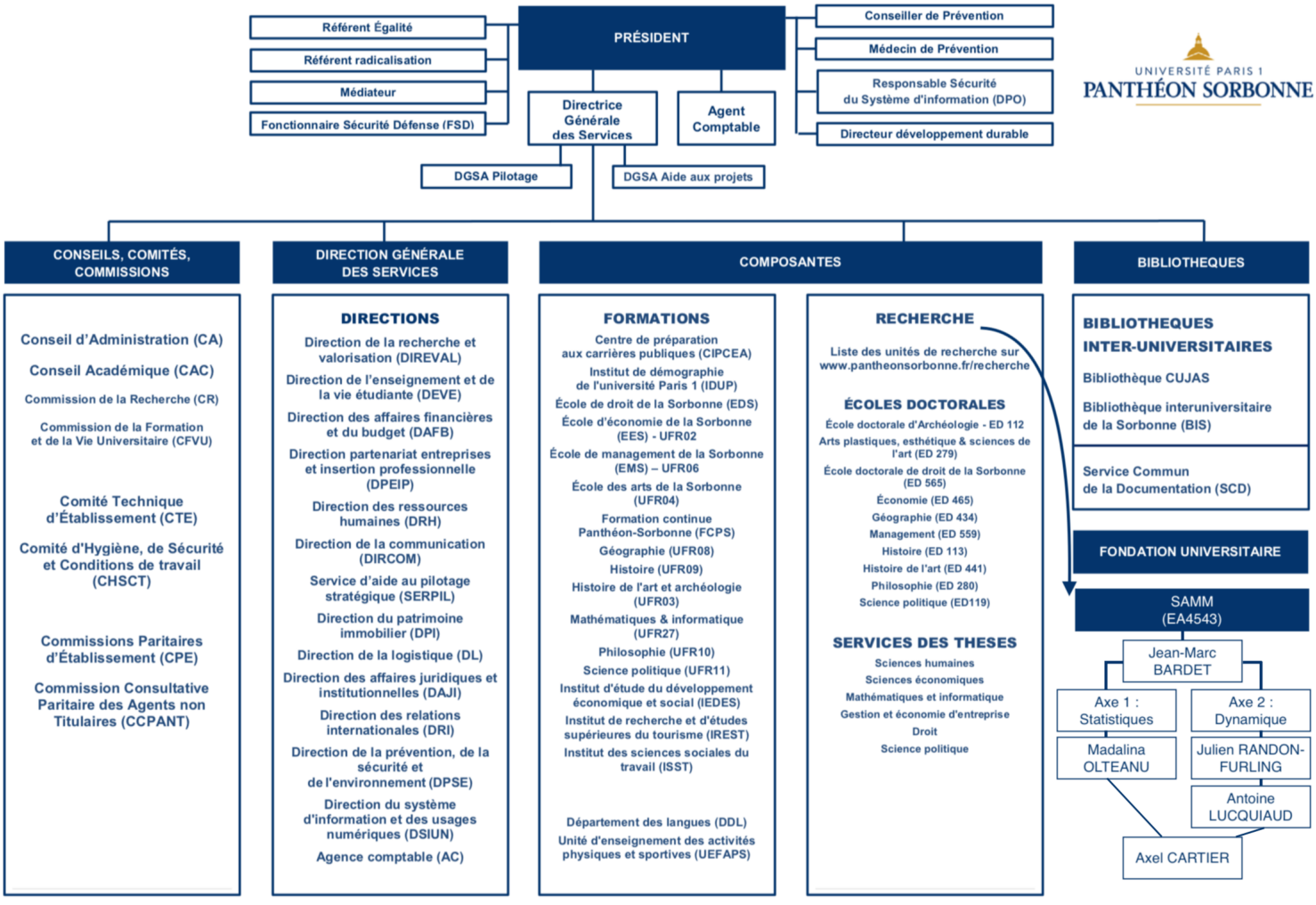
U

UNIMED Universités de Méditerranée. 5

UT Unités de Traitement. 8

Annexes

I. Organigramme



II. Programme R

```r
FD_LOGEMTZA_2013 <- read.csv("/Users/Maxwel/Desktop/WORK EN COURS/STAGE/DATA/rp2013_logemtza_txt/FD_LOGEMTZA_2013
DATA<- FD_LOGEMTZA_2013[which(FD_LOGEMTZA_2013$COMMUNE==75056),]

DATA<- DATA[which(DATA$CATL==1),c(3,28)]

DF2<-DATA %>% group_by(IRIS) %>% count(HLML)


CLOGS=DF2[DF2$HLML==1,-2]
CLOGT= aggregate(DF2$n, by=list(Category=DF2$IRIS), FUN=sum)
colnames(CLOGT)[1] <- "IRIS"

DFF<-merge(CLOGS,CLOGT,by="IRIS",all.y = TRUE)
colnames(DFF)[1] <- "CODE_IRIS"
colnames(DFF)[2] <- "logs"
colnames(DFF)[3] <- "logt"
DFF[is.na(DFF)] <- 0

write.table(DFF, file = "/Users/Maxwel/Desktop/WORK EN COURS/STAGE/LOGPARISF.CSV", sep = ";")
```

III. Extraits des bases de données

III.1 Données brutes

FD_LOGEMTZA_2013.txt [9] Ici, chaque observation représente un ménage enquêté par l'INSEE. Les poids (IPONDL) permettent de se ramener à l'effectif réel estimé dans l'IRIS et la variable HLML indique si un logement appartient à un organisme HLM ou non.

COMMUNE	ARM	IRIS	...	HLML	...	IPONDL	...	TRIRIS	TYPC	TYPL	VOIT	WC
75056	75101	751010101	...	2	...	3.586660	...	750011	3	2	1	Z
75056	75101	751010101	...	2	...	3.586660	...	750011	3	2	1	Z
75056	75101	751010101	...	2	...	3.586660	...	750011	3	2	0	Z
75056	75101	751010101	...	2	...	3.586660	...	750011	3	2	1	Z
75056	75101	751010101	...	2	...	3.586660	...	750011	3	2	1	Z
75056	75101	751010101	...	2	...	1.302288	...	750011	3	2	X	Z
75056	75101	751010101	...	2	...	3.544388	...	750011	3	2	0	Z
75056	75101	751010101	...	2	...	3.740054	...	750011	3	2	0	Z
75056	75101	751010101	...	2	...	1.302286	...	750011	3	2	0	Z
75056	75101	751010101	...	1	...	3.740059	...	750011	3	2	X	Z
...	...	...	...	...	...	...	...	...	...	...	...	...

CONTOURS-IRIS.shp [10] C'est le fichier de forme contenant tous les IRIS de la métropole française : une observation représente donc un IRIS.

	INSEE_COM	NOM_COM	IRIS	CODE_IRIS	NOM_IRIS	TYP_IRIS	geometry
0	01001	L'Abergement-Clémenciat	0000	010010000	L'Abergement-Clémenciat	Z	POLYGON ((846847.6999999131 6563787.000003602,...
1	01002	L'Abergement-de-Varey	0000	010020000	L'Abergement-de-Varey	Z	POLYGON ((889088.1999998859 6549317.300003688,...
2	01004	Ambérieu-en-Bugey	0102	010040102	Longeray-Gare	H	POLYGON ((882210.5999998887 6542241.900003731,...
3	01004	Ambérieu-en-Bugey	0202	010040202	Tiret-Les Allymes	H	POLYGON ((882210.5999998887 6542241.900003731,...
4	01004	Ambérieu-en-Bugey	0101	010040101	Les Perouses-Triangle d'Activite	H	POLYGON ((880308.3099998899 6542546.350003731,...
...	...	...	...	...	...	...	...
50148	69255	Vaugneray	0000	692550000	Vaugneray	Z	POLYGON ((825415.0999999206 6513888.200003888,...
50149	71578	Clux-Villeneuve	0000	715780000	Clux-Villeneuve	Z	POLYGON ((865238.2999999148 6651468.600003094,...
50150	72137	Villeneuve-en-Perseigne	0000	721370000	Villeneuve-en-Perseigne	Z	POLYGON ((497663.2000000729 6823406.000002102,...
50151	73263	Saint-Offenge	0000	732630000	Saint-Offenge	Z	POLYGON ((933179.5999998527 6518449.700003878,...
50152	77316	Orvanne	0000	773160000	Orvanne	Z	POLYGON ((685093.8000000056 6805315.100002186,...

50153 rows × 7 columns

III.2 Tables intermédiaires (3.3.2)

DataFrame PARIS Cette base de données provient du shapefile et ne contient que les IRIS qui nous intéressent pour notre étude.

	INSEE_COM	NOM_COM	IRIS	CODE_IRIS	NOM_IRIS	TYP_IRIS	geometry	x	y
39125	75101	Paris 1er Arrondissement	0204	751010204	Les Halles 4	H	POLYGON ((652285.8100000154 6862819.570001863,...	652040.040864	6.862820e+06
39126	75101	Paris 1er Arrondissement	0206	751010206	Les Halles 6	A	POLYGON ((651617.7000000156 6862947.100001859,...	651763.426941	6.862942e+06
39127	75101	Paris 1er Arrondissement	0101	751010101	Saint-Germain l'Auxerrois 1	H	POLYGON ((652186.3000000154 6862282.900001862,...	651972.333344	6.862249e+06
39128	75101	Paris 1er Arrondissement	0202	751010202	Les Halles 2	H	POLYGON ((651540.0000000156 6862651.000001862,...	651825.198123	6.862489e+06
39129	75101	Paris 1er Arrondissement	0102	751010102	Saint-Germain l'Auxerrois 2	A	POLYGON ((651923.0600000157 6861773.040001867,...	651868.072429	6.861968e+06
...	...	...	...	...	...	...	...	...	...
40112	75120	Paris 20e Arrondissement	8020	751208020	Charonne 20	H	POLYGON ((656719.000000014 6861954.000001867, ...	656752.088998	6.862242e+06
40113	75120	Paris 20e Arrondissement	8004	751208004	Charonne 4	H	POLYGON ((656140.0000000142 6861270.000001867,...	656021.722969	6.861246e+06
40114	75120	Paris 20e Arrondissement	7803	751207803	Saint-Fargeau 3	H	POLYGON ((656644.000000014 6863260.00000186, 6...	656717.385549	6.863247e+06
40115	75120	Paris 20e Arrondissement	7911	751207911	Pere Lachaise 11	H	POLYGON ((655470.1000000143 6863133.00000186, ...	655385.264316	6.863319e+06
40116	75120	Paris 20e Arrondissement	7913	751207913	Pere Lachaise 13	H	POLYGON ((655860.0000000142 6862956.00000186, ...	655716.551187	6.863043e+06

987 rows × 9 columns

LOGPARISF2015.CSV C'est la table qui compte le nombre de logements et de logements sociaux par IRIS. Elle comporte autant d'observations que la table finale soit 942.

CODE_IRIS	logs	logt
751010101	22	141
751010102	0	28
751010103	8	73
751010104	0	5
751010201	24	606
751010202	4	343
751010203	40	654
751010204	145	469
751010206	0	143
751010301	8	574
...	...	...

III.3 Table Finale (3.3.3)

	x	y	CODE_IRIS	NOM_IRIS	geometry	logs	logt	taux_base	sorted_index	state_logt	state_logs	taux	check
0	650674.985472	6.861731e+06	751072501	Saint-Thomas d'Aquin 1	POLYGON ((650515.0000000158 6861973.000001865,...	14.521736	1085.346712	0.013380	[[0, 369, 148, 783, 785, 268, 602, 134, 922, 1...	[[1085.34671236944, 1933.098138685339, 3196.04...	[[14.5217355925758, 21.89037105602724, 28.2763...	[[0.013379812577008814, 0.011323983308428634, ...	918
1	653937.006791	6.863163e+06	751114108	Folie Mericourt 8	POLYGON ((654039.8000000148 6863141.800001861,...	30.149543	1392.288841	0.021655	[[1, 915, 766, 890, 547, 806, 549, 112, 188, 8...	[[1392.2888406442598, 2351.5121917947968, 3428...	[[30.149542997140102, 30.149542997140102, 175....	[[0.021654661099767828, 0.012821342412062257, ...	306
2	648386.749492	6.861844e+06	751072807	Gros Caillou 7	POLYGON ((648699.7200000165 6861458.710001869,...	3.058440	1082.686461	0.002825	[[2, 105, 663, 642, 305, 136, 42, 174, 32, 469...	[[1082.68646092197, 2715.06852233536, 3698.655...	[[3.058440466666768, 146.9818451319357, 305.072...	[[0.002824862577539985, 0.054135593235602616, ...	920
3	654131.988363	6.859125e+06	751135015	Gare 15	(POLYGON ((653788.0000000149 6859452.000001879...	77.545073	340.417197	0.227794	[[3, 769, 296, 763, 598, 191, 119, 63, 195, 69...	[[340.417196656083, 2595.254905977783, 3042.51...	[[77.5450732476938, 493.02344774239975, 510.97...	[[0.22779423016645103, 0.18997110711814616, 0....	1
4	655814.799115	6.863656e+06	751207815	Saint-Fargeau 15	POLYGON ((655936.0000000142 6863800.800001855,...	146.064460	1510.050754	0.096728	[[4, 433, 705, 353, 441, 206, 27, 215, 262, 57...	[[1510.05075358772, 3033.67417294514, 4295.483...	[[146.0644602282, 781.5666445040059, 1165.6435...	[[0.09672817942123228, 0.25763038479021905, 0....	600
...	...	...	...	...	...	...	...	...	...	...	...	...	...
937	655957.187471	6.861627e+06	751208015	Charonne 15	POLYGON ((655765.7500000142 6861588.020001866,...	206.831391	1412.311041	0.146449	[[937, 224, 730, 525, 814, 405, 378, 526, 774,...	[[1412.3110406219698, 2899.9296959522, 4283.74...	[[206.83139136735699, 529.889088859921, 708.03...	[[0.1464488950509586, 0.1827248052252972, 0.16...	1
938	648754.205899	6.865501e+06	751176612	Plaine Monceau 12	POLYGON ((648651.0000000163 6865564.000001845,...	5.696138	1579.590313	0.003606	[[938, 905, 749, 239, 907, 847, 599, 210, 504,...	[[1579.5903134170599, 3378.35383939297, 4514.6...	[[5.69613788413581, 1726.435751700296, 1751.32...	[[0.0036060856006476768, 0.5110286943804874, 0...	26
939	654980.398010	6.863832e+06	751207711	Belleville 11	POLYGON ((654745.2500000144 6863896.160001854,...	1135.857066	1902.809538	0.596937	[[939, 650, 579, 883, 640, 371, 86, 38, 385, 5...	[[1902.8095375415803, 3203.89411757378, 4631.5...	[[1135.85706560571, 1636.776508124702, 1935.80...	[[0.5969368153752431, 0.5108709739022798, 0.41...	463
940	654116.435627	6.863729e+06	751104001	Hopital Saint-Louis 1	POLYGON ((654185.2300000146 6863881.680001854,...	41.738755	1380.426960	0.030236	[[940, 892, 633, 884, 588, 346, 282, 547, 351,...	[[1380.4269596043698, 2907.12588083009, 3793.9...	[[41.738754608623296, 651.3419413112692, 854.9...	[[0.030236119570270937, 0.22405013336584087, 0...	4
941	646584.145251	6.863264e+06	751166305	Porte Dauphine 5	POLYGON ((646249.1000000173 6862874.90000186, ...	26.343508	910.536162	0.028932	[[941, 576, 208, 379, 5, 646, 462, 619, 900, 7...	[[910.5361618097161, 2561.537799219376, 3699.4...	[[26.3435076964472, 26.3435076964472, 26.34350...	[[0.02893186322670452, 0.010284254912996144, 0...	911

942 rows × 13 columns

IV. Planning prévisionnel et cahier des charges

IV.1 Planning prévisionnel

Période concernée	Objectifs
08-04 au 14-04	Lecture de l'article et compréhension de la méthode
15-04 au 12-05	Choix de l'environnement technique et recherches
12-05 au 19-05	Entraînement sur une grille
19-05 au 26-05	Reproduction des résultats
26-05 au 17-06	Rédaction du rapport et préparation de la soutenance
17-06 au 28-06	Prolongements et autres applications

IV.2 Cahier des charges

Présentation du projet Dans le cadre d'un stage au SAMM, il m'est demandé de programmer une nouvelle méthode d'analyse spatiale développée par le laboratoire dans un langage de programmation libre d'accès et gratuit. Il faut ainsi que je retrouve les résultats de l'article scientifique dans lequel cette méthode a été utilisée.

Equipe et matériel Je dois réaliser cette mission seul même si je suis ponctuellement aidé par les collaborateurs du laboratoire. Pour cela, je suis équipé de mon ordinateur personnel.

Budget Aucune enveloppe n'a été dégagée pour ce projet.

Délais La date maximale est fixée au 27 mai pour la remise du livrable.

Contraintes

Complexité en temps Ce programme doit fournir des résultats dans un temps acceptable à l'échelle humaine, soit moins de 20 minutes pour les données parisiennes.

Réutilisabilité du programme La méthode d'agrégation doit pouvoir s'appliquer à différents cadres, il est important que le programme soit facilement réutilisable sur d'autres bases de données.

V. Programme Python

```
from sklearn.neighbors import KDTree
import numpy as np
import pandas as pd
from sklearn.utils import shuffle
import matplotlib.pyplot as plt
import datetime
import geopandas as gp
## RECUPERATION TEMPS ##
time=datetime.datetime.now()
## IMPORT + PREPARATION DATA ##
data = pd.read_csv("/Users/Maxwel/Desktop/WORK EN COURS/STAGE/
    LOGPARISPF.CSV", sep=";")
IRIS_B=gp.read_file('/Users/Maxwel/Desktop/WORK EN COURS/STAGE/DATA/
    CONTOURS-IRIS_2-1__SHP_LAMB93_FXX_2016-11-10/CONTOURS-IRIS/1
    _DONNEES_LIVRAISON_2015/CONTOURS-IRIS_2-1_SHP_LAMB93_FE-2015/
    CONTOURS-IRIS.shp')
PARIS=IRIS_B[IRIS_B['NOM_COM'].str.match('Paris')]
PARIS=PARIS.drop(PARIS[PARIS["NOM_COM"]=="Parisot"].index)
PARIS=PARIS.drop(PARIS[PARIS["NOM_COM"]=="Paris-l'Hôpital"].index)
PARIS=PARIS.drop(PARIS[PARIS['NOM_IRIS'].str.match('Bois')].index)
PARIS=PARIS.drop(PARIS[PARIS['NOM_IRIS'].str.match('Jardin du
    Luxembourg')].index)
PARIS['x']=PARIS.geometry.centroid.x
PARIS['y']=PARIS.geometry.centroid.y
df=PARIS.ix[:,('x','y','CODE_IRIS','NOM_IRIS','geometry')]
df['CODE_IRIS']=pd.to_numeric(df.CODE_IRIS)
data2= pd.merge(df,data,on="CODE_IRIS",how="outer")
data2=data2.fillna(value=0)
data2=data2[data2["logt"]!=0]
df=data2
df['taux_base']=df["logs"]/df["logt"]
df['sorted_index']= None
df['state_logt']= None
df['state_logs']= None
df['taux']= None
df['check']=None
## RANDOMISATION INDICES ##
df=shuffle(df)
df=df.reset_index(drop=True)
## CALCUL DES SEUILS ##
taux_global=df.logs.sum()/df.logt.sum()
seuil=0.05
```

```
s= [taux_global*(1-seuil), taux_global*(1+seuil)]
## CALCUL KDTREE ##
kdt = KDTree(df[['x', 'y']])
## CALCUL INDEX PLUS PROCHES VOISINS ##
for x in range(len(df)):
    origin= np.stack((df.x[x],df.y[x]),axis=-1)
    origin_kdt = np.expand_dims(origin, axis=0)
    nearest_point_index = kdt.query(origin_kdt, k=len(df),
    return_distance=True)
    df.sorted_index[x]=nearest_point_index[1]
    df.state_logs[x]= np.ones((1,len(df)))
    df.state_logt[x]= np.ones((1,len(df)))
    df.taux[x]= np.ones((1,len(df)))
## CALCUL DES NIVEAUX D'AGREGATION ##
for i in range(len(df)) :
    for j in range(len(df)) :
        if j == 0 :
            df.state_logt[i][0,j]=df.logt[i]
            df.state_logs[i][0,j]=df.logs[i]
            df.taux[i][0,j]=df.state_logs[i][0,j]/df.state_logt[i][0,
    j]
            if s[0] <df.taux[i][0,j]<s[1] :
                df.check[i]= 0
        else :
            df.state_logt[i][0,j]= df.state_logt[i][0,(j-1)] + df.
    logt[df.sorted_index[i][0,j]]
            df.state_logs[i][0,j]= df.state_logs[i][0,(j-1)] + df.
    logs[df.sorted_index[i][0,j]]
            df.taux[i][0,j]= df.state_logs[i][0,j]/df.state_logt[i
    ][0,j]
            if (s[0] <df.taux[i][0,j]<s[1]) and (df.check[i] is None)
     :
                df.check[i]=j
## AFFICHAGE ##
pd.set_option("display.max_columns", df.shape[1])
pd.set_option("display.max_rows", df.shape[0])
df
## GRAPHIQUES ##
plt.figure()
plt.title("Trajectoires")
plt.axhline(y=s[0],color="black")
plt.axhline(y=s[1],color="black")
for i in range(len(df)) :
    plt.plot(pd.Series(df.taux[i].ravel()))
    plt.axvline(x=df.check[i])
plt.show()
df.plot(column='check',cmap='plasma')
df.plot(column='taux_base',cmap='viridis')
```

```
80 ## CALCUL DU TEMPS DE TRAVAIL ## 15-23 MIN POUR 992 IRIS DE PARIS
       (5-13MIN POUR 942)
   datetime.datetime.now()-time
```