

Projet d'Études Entrants 2017

Projet d'études N° 35

Noms des élèves :

BERTRAND Theo
CHALICARNE Raphaël
GALLIMARD Maude
LI Qifei
LYON Fabien
RADISIC Katarina

Commanditaires :

DERRODE Stéphane
VUILLEMOT Romain

Tuteur(s) scientifique(s) :

DERRODE Stéphane
VUILLEMOT Romain

Conseiller en communication :

HOURCADE Nicolas

Conseiller en gestion de projet :

GIAUQUE Alexis

Département d'accueil :

Mathématiques - Informatique - M.I.

Date du rapport :

8 Juin 2018

**Collecte et analyse des données GPS
sur la mobilité des étudiants de l'ECL**

Résumé

L'étude de la mobilité des usagers sur le campus est un problème intéressant afin de pouvoir prévoir un agencement efficace du campus de l'Ecole Centrale de Lyon. Cette étude de la mobilité des acteurs sur le campus est orientée vers une réflexion sur l'acquisition de données de géolocalisation, la réalisation à petite et à petite-moyenne échelle d'une telle acquisition et propose quelques essais pour tirer des informations d'une telle collecte. Ainsi, cette analyse permet d'obtenir quelques résultats qualitatifs afin de comparer la réalité aux modèles, correspondant à des présupposés sur la vie du campus.

Au cours de l'étude, une démarche d'essais pas à pas est mise en place afin d'adopter à chaque étape un regard critique sur les résultats et de préparer les étapes suivantes. La méthodologie adoptée sera développée précisément dans ce rapport.

Ainsi, un moyen simple d'obtenir les données de localisation est sélectionné, une collecte est menée auprès de quelques élèves de l'école, puis la qualité des données récoltées est estimée, tout en tirant quelques résultats qualitatifs.

Abstract

The study of users' mobility on-campus is an interesting problem whose goal is to efficiently organise the campus of the Ecole Centrale de Lyon. This study has been conceived as a reflexion on geolocation data acquisition, the realisation of such acquisition on a small and small-medium scale and the visualisation of this data, all of this in order to extract relevant information. The analysis of this data yields the results which enable the comparison between the models adopted and reality.

The approach taken to this research is a step-by-step method, the reason being the need of staying critical about the results and preparing the further stages of the project. The methodology undertaken in the study will be explained exhaustively in this report.

Thus, a simple way to collect location data is chosen, furtherly a collect with other ECL students is carried out, and in the end the quality of collected data is estimated while trying to maximise the information obtained.

Remerciements

Nous tenons à remercier ici toutes les personnes qui ont contribué d'une manière ou d'une autre à la réalisation de notre projet tout au long de cette année.

Nous souhaitons avant tout remercier MM. Stephane DERRODE et Romain VUILLEMOT qui ont chacun assumé la double casquette de commanditaire et de tuteur pour nous proposer ce projet et nous aider à le mener à bien. Nous leur sommes reconnaissants pour la pertinence de leurs commentaires au cours des différentes rencontres et pour nous avoir guidé afin d'ajuster les objectifs du projet.

Nous adressons également nos remerciements à M. Nicolas HOURCADE, notre conseiller en communication, pour la qualité de ses conseils au fil des réunions, des conférences et des travaux dirigés du mercredi après-midi. Il en va de même pour M. Alexis GIAUQUE, notre conseiller en gestion de projet, dont les interventions nous ont permis de mieux saisir les enjeux relatifs à l'organisation du projet, désormais le concept de planning prévisionnel ne nous échappera plus.

Nous remercions sincèrement l'ensemble des étudiants qui ont acceptés de participer à notre projet en collectant leurs propres données GPS et en nous faisant parvenir l'historique de leurs positions. Un grand merci à Bob AUBOUIN, Yue BAI, Ahmed BELHOUARI, Hamza BENADADA, Bob BETY, Matthieu BORDEAU, Rémi CECCHINATO, Etienne CHAVASSE FRETAZ, Maria CHERRADI, Jules COLAS, Geoffroy DE BEAUMONT, Erwan DE LONGCAMP, Augustin DURIVAUT, Guillaume FREGOSI, Agathe GAMBERT, Corentin HANSER, Théo KAPRELIAN, Thomas MAURAS, Mehdi MOBAREK, Bilel OMRANI, Pauline PERRIN, Céline RACQUE, Vijay TIROU, Nan WU, Tong WU et Zhaozhi WU pour leur implication, avec une mention spéciale pour Ahmed BELHOUARI et Jules COLAS qui nous ont prêté leurs téléphones portables par la suite et nous ont du même pas accompagnés pendant nos mesures de terrain. Nous remercions au même titre Hugo SAPIN et Martina CARPANI pour leur joyeuse compagnie lors des trajets réalisés pour les expériences.

Table des matières

Résumé	2
Remerciements	3
Table des matières	4
Table des figures et des tableaux	6
Introduction	7
1. Contexte scientifique et technique	8
1.1 Contexte	8
1.2 Méthodologie retenue	8
2. Présentation des aspects théoriques	9
2.1 Etude Moovendi	9
2.2 Positionnement GPS	9
3. Démarche suivie	11
3.1 Choix de l'application pour la collecte des données GPS	12
3.2 Analyse des paramètres influençant la qualité des données GPS	15
3.2.1 Démarche expérimentale	15
3.2.2 Configuration optimale de Google Maps	16
3.2.3 Paramètres considérés et hypothèses de départ	17
3.2.4 Protocole pour chaque expérience	18
3.3 Démarche informatique	19
3.3.1 Définition de zones d'intérêt sur le campus	19
3.3.2 Application de la méthode K-means aux données de la collecte	19
3.3.3 Visualisation cartographique des données collectées	22
3.3.4 Etablissement d'un graphe de déplacements entre les zones du campus	24
4. Résultats obtenus	26
4.1 Résultats de l'analyse des paramètres influençant la qualité de la collecte	26
4.1.1 Influence des paramètres extérieurs	26
4.1.1.1 Indépendance des jours de la semaine	26
4.1.1.2 Indépendance des créneaux horaires	27
4.1.1.3 Influence de la météo	28
4.1.2 Influences des paramètres liés à l'application Google Maps	29
4.1.2.1 Influence du nombre des autres applications actives	29
4.1.2.2 Influence des autres applications de trackage	30
4.1.3 Rôle des paramètres internet	31
4.1.4 Influence des personnes alentours	34

4.1.5 Influence du niveau de batterie	34
4.1.6 Influence du téléphone en lui-même : modèle, Android/ iOS, version	34
Résumé des résultats obtenus	34
4.2 Présentation des différentes visualisation obtenues	37
4.2.1 Zones d'intérêt issues de la méthode K-means	37
4.2.2 Représentations cartographiques	37
4.2.3 Graphe de déplacements entre les zones d'intérêt	41
Conclusion	44
Bibliographie	45
Annexes techniques	47
Check-list de rapport de Projet d'Études	48

Table des figures et des tableaux

Figures

1.	Trajet obtenu avec Google Maps	16
2.	Trajet (1) en orange avec les points définies en bleu	18
3.	Division du campus en zones géographiques	19
4.	Ensemble des points collectés auprès des participants	21
5.	Initialisation de la méthode des K-means	22
6.	Résultat final de la méthode pour ces conditions initiales	22
7.	Mesures obtenues avec les mêmes paramètres le Mercredi 16/05/2018	26
8.	Mesures obtenues avec les mêmes paramètres le Vendredi 18/05/2018	26
9.	Trajet (1) en orange avec les points définies en bleu	26
10.	Exemple de deux trajets faits le lundi 21/05/2018, le premier entre 13h40-13h54 et le deuxième entre 15h44-16h00	27
11.	Tracé GPS pour étude de l'influence des applications ouvertes en parallèle	29
12.	Trajet habituel	30
13.	Trajet réalisé avec FollowMee active	30
14.	Cercle d'incertitude de position dans le bâtiment X	31
15.	Cercle d'incertitude dans le bâtiment X, proche d'une fenêtre	31
16.	Cercle d'incertitude dans le bâtiment X, proche d'une fenêtre après temps de transition	31
17.	Trajet dans le W1bis	32
18.	Trajet dans le W1	32
19.	Tracé GPS du 06/06 entre 16h et 18h pour étude du rôle d'internet	33
20.	Ensemble des points récupérés pendant la collecte, une couleur par participant	37
21.	Nombre de points par participants	38
22.	Heatmap des données de tous les jours pour toutes les heures	29
23.	Heatmaps montrant la densité, le jeu de points ayant été découpé par tranches de 2h, de 8h à 20h, dans l'ordre de gauche à droite et de haut en bas	41
24.	Graphe de transition pour les données accumulées sur la journée	42
25.	Graphes des transitions de 7h à 12h et de 13h à 17h	43

Tableaux

1.	Avantages et inconvénients de différentes applications de géolocalisation	13
2.	Paramètres influençant la qualité des données acquises et hypothèses sur leur influence	17
3.	Résumé des résultats obtenus avec l'analyse des paramètres	36
4.	Les différents partitionnements obtenus avec la méthode k-means sur l'ensemble des données collectées	37

Introduction

La connaissance des flux de populations est aujourd’hui en pleine expansion. L’obtention de telles informations présente des intérêts commerciaux, sociaux, économiques et politiques, ce qui explique l’engouement pour ce domaine de recherche. L’étude à grande échelle des mouvements de population est désormais permise par la récente démocratisation d’appareils mobiles connectés équipés pour la plupart de systèmes de géolocalisation. En parallèle le développement des ressources faisant appel à l’intelligence collective sur le modèle des outils open-source donne accès rapidement à des informations précises sur les territoires étudiées.

A l’échelle de l’Ecole Centrale de Lyon, les différentes zones géographiques du campus font face à des périodes d’affluence et de creux avec une incidence sur la fluidité du trafic, la gestion de la consommation énergétique, les contacts humains ou encore la gestion des interventions de secours et de manutention. Une connaissance spatio-temporelle de ses afflux pourrait ainsi permettre une optimisation de l’aménagement du campus et de son fonctionnement.

Fort de ces constatations, ce projet s’intéresse à l’étude de la mobilité des étudiants sur le campus de l’Ecole Centrale de Lyon. Il a vocation d’une part à appréhender les techniques de visualisation de données GPS pour une petite population de manière à s’affranchir des difficultés liées au stockage de données et au temps de calcul. D’autre part, il a pour ambition d’étudier les contraintes liées à la précision de la géolocalisation à une petite échelle géographique telle que celle du campus.

Ce projet est donc destiné à comprendre comment caractériser la dynamique des mouvements sur le campus de l’Ecole Centrale de Lyon en faisant face aux limites liées à l’acquisition des données GPS des étudiants à cette échelle. Il s’agit de collecter des données GPS, d’étudier la qualité de ces données au niveau du campus de l’ECL et de cartographier et visualiser de manière numérique la dynamique des déplacements et fréquentations des diverses zones du campus.

Pour ce faire, des hypothèses sont préalablement établies à partir de connaissances scientifiques et de conjectures intuitives. Les résultats obtenus par les expériences et les simulations numériques viennent alors en confirmation ou en contradiction de ces hypothèses. Deux types d’acquisitions de données sont ainsi réalisées : des collectes de données personnelles et une collecte d’une semaine sur un panel d’une vingtaine d’étudiants. Pour faciliter l’étude numérique des flux, le campus est divisé en plusieurs zones géographiques appelées zones d’intérêt.

Ce document présente plus en détails les enjeux de ce projet ainsi que la méthodologie adoptée. Il fait l’état des connaissances générales qui justifient les décisions adoptées. Il met en évidence les démarches menées et les résultats obtenus concernant l’acquisition des données de géolocalisation, l’analyse des paramètres influençant la qualité des données de géolocalisation et la visualisation numérique des densités et des flux d’étudiants sur le campus de l’Ecole Centrale de Lyon.

1. Contexte scientifique et technique

1.1 Contexte

Ce projet d'études (PE) a deux objectifs principaux. Le premier consiste à mettre en œuvre une démarche d'analyse de la qualité des données GPS pouvant être acquises sur le campus de l'Ecole Centrale de Lyon (ECL). Le second, découlant des résultats du premier, est de comprendre l'organisation des déplacements sur le campus à partir des données récoltées.

Ce projet a pour ambition de déterminer les paramètres pouvant jouer sur la qualité des données GPS récoltées et leur influence. Il a aussi pour but d'évaluer la fréquentation des principales zones du campus pendant la semaine aux différentes heures de la journée grâce à des graphes de transition entre zones.

Face à cela, des problèmes de stockage de données, d'anonymisation peuvent se poser. En effet, des données précises de géolocalisation sont récoltées mais celles-ci doivent rester privées, c'est-à-dire anonymes, pour des questions de sécurité et de vie privée. Il y a donc derrière ce projet un problème de Big Data et donc des enjeux autres que la technique. L'étude menée s'affranchit toutefois de ces problèmes, car les collectes de données seront réalisées sur un petit effectif, dont les individus sont familiers des membres de l'équipe.

Les enjeux liés à la compréhension des déplacements sur le campus et aux conclusions de ce projet pourront à terme être multiples. Adapter l'éclairage du campus ou le chauffage des bâtiments en fonction de leur fréquentation ou encore modifier les emplois du temps pour éviter une surfréquentation du restaurant universitaire le midi sont autant de perspectives qui pourront être explorées par la suite.

1.2 Méthodologie retenue

1.2.1 Aspect analyse de la qualité des données GPS

Pour mener à bien cette étude, une méthodologie de "benchmark" est adoptée. L'idée principale est de formuler une hypothèse pour chaque paramètre pouvant influencer la qualité des données GPS récoltées (météo, heure de la journée, Internet activé...). On vérifie ensuite cette hypothèse en empruntant un trajet de référence de sorte à mettre en évidence l'influence du seul paramètre testé. Par exemple, on parcourt le trajet de référence par beau temps et par temps couvert, sans changer aucun autre paramètre, et on regarde si les trajectoires sont modifiées ou non par la météo.

1.2.2 Aspect analyse des déplacements sur le campus

Afin de disposer de données brutes sur lesquelles analyser les déplacements des Centraliens, un panel d'étudiants est recruté et chacun mesure ses trajectoires sur le campus durant une semaine (du Jeudi 1/03/18 au Vendredi 9/03/18). Le choix a été fait de ne tenir compte d'aucune information pouvant discriminer les candidats, c'est-à-dire que la promotion, le sexe ou la nationalité des participants ne sont pas relevés, bien que ces critères pourraient être utilisés pour étudier les trajectoires privilégiées empruntées par les participants.

Chacun transmet ensuite l'historique de ses trajectoires au format *json* en le récupérant sur Google Takeout et en le déposant sur un dossier de la plateforme ECloud où seuls les membres du projet disposent d'un accès. Ces données ont ensuite pu être analysées afin de déterminer quelles sont les zones du campus les plus fréquentées et à quelles heures.

Pour faciliter l'analyse des données obtenues, le campus a été découpé en différentes zones géographiques qui représentent les lieux emblématiques de la vie étudiante du campus (résidences, W1/W1b, M16 et RU...). Deux techniques ont été utilisées pour découper le campus.

Une approche empirique est choisie en découplant intuitivement le campus en différentes zones. Un algorithme (K-means [1]) est aussi utilisé pour vérifier que le découpage est acceptable en terme de nombre et positions des différentes zones.

L'analyse développée dans cette étude s'inspire de travaux d'ores et déjà menés sur les déplacements à l'échelle du campus. Elle tient compte des informations ayant pu en être extraites.

2. Présentation des aspects théoriques

2.1 Etude Moovendi

Lors de l'année 2016-2017, des étudiants de troisième année de la filière ADE suivant l'option MD ont réalisé une mission dans laquelle ils discutaient de la possibilité d'installer une plateforme de mobilité sur le Campus Lyon Ouest Ecully. Dans le cadre de leur mission, ils ont déterminé les moyens de transports utilisés par les étudiants et personnels pour se rendre au campus, et dans quelles proportions. Ils ont également déterminé à quels horaires les gens entraient et quittaient le campus.

D'après eux, beaucoup d'étudiants arrivent avant 8h du matin pour les cours, puis les gens quittent le campus de manière éparse l'après-midi. Il s'agit donc d'une hypothèse qu'il faudra vérifier. [2]

2.2 Positionnement GPS

GPS (Global Positioning System) est le système de géolocalisation le plus fréquemment utilisé, et fonctionne avec une constellation de 30 satellites en orbite autour de la Terre. Chaque satellite envoie sur Terre des signaux sous forme d'ondes électromagnétiques qui portent des informations : sa position dans l'espace et l'heure et la date d'émission du signal [3].

Les satellites disposent de leurs éphémérides, c'est à dire qu'ils connaissent leurs propres positions sur leur orbite. Ils savent, par exemple, à quelle heure et quel jour ils vont passer au dessus de Paris. Tout ceci est calculé à l'avance avec une grande précision. Ces données sont parfois mises à jour. Ces éphémérides sont envoyées au capteur GPS (par exemple celui du téléphone) dans le signal GPS. Le capteur dispose alors de la trajectoire du satellite pour les 4 heures qui viennent. Un appareil utilise la méthode "trilateration" pour calculer sa propre latitude, longitude et altitude, et donc donner sa position [4].

Habituellement, les antennes GPS qui sont dans un smartphone ne sont pas aussi puissantes car l'espace intérieur au sein des smartphone est assez limité. L'inconvénient est l'incapacité de l'utiliser en intérieur et le fait que ça va prendre beaucoup de temps pour recevoir toutes les données. En fait, tout blocage (par exemple le bâtiment) va dégrader le signal GPS, le signal sera donc considérablement atténué aussi bien à l'intérieur qu'à l'extérieur, c'est la raison pour laquelle on utilise d'autres techniques de géolocalisation pour les espaces intérieurs [5].

On peut constater qu'il y a plusieurs inconvénients à l'utilisation d'une seule technique de géolocalisation. Le positionnement est meilleur lorsqu'on les combine, c'est le cas pour la plupart des smartphones. D'abord "A-GPS", dont le système est presque le même que celui de GPS, sauf que les éphémérides passent par le réseau Internet (Réseaux mobiles ou Wifi), est un système largement utilisé par la plupart des appareils. L'avantage est qu'il améliore la vitesse de localisation à l'aide du réseau Internet : le signal GPS se fait à 50 octets par seconde, alors que la 4G monte à 50 millions d'octets par seconde. Ensuite, la

géolocalisation par GSM permet le positionnement d'un terminal GSM en se basant sur certaines informations relatives aux antennes GSM auxquelles le terminal est connecté. Son avantage est que l'on peut se localiser sans WIFI rapidement. Enfin, la géolocalisation par WIFI utilise la même méthode qu'un terminal GSM en se basant sur les identifiants des bornes Wi-Fi qu'il détecte. L'avantage est que ce type de géolocalisation est rapide et assez précis quand il y a beaucoup de monde alentour. Son inconvénient est qu'il faut se connecter à la WIFI pour obtenir l'information [3].

Avec la combinaison de ces techniques, le temps de réponse à l'allumage et l'adaptabilité sont améliorés. Cela permet par exemple de localiser des gens en intérieur en utilisant le WIFI ou le GSM et de localiser des gens à l'extérieur à l'aide du GPS. L'iPhone d'Apple est un exemple de terminal capable d'utiliser une méthode de géolocalisation grâce à GSM/Wi-Fi/A-GPS/GPS [3].

Des sources d'erreurs peuvent s'introduire dans chacun des quatre domaines. D'abord la réflexion du signal sur des objets proches du récepteur produit des échos qui interfèrent parfois sur le signal reçu et provoquent ainsi un décalage. Les signaux de satellites peuvent également être réfléchis par des surfaces de vêtement voire même des murs : c'est le problème des multi-trajets. Dans ce cas le géonavigateur ne réceptionne que des échos des signaux et la géolocalisation calculée tarde à se stabiliser. Il est important de noter que l'erreur typique ne considère pas les effets dus à des réflexions parasites supérieures à un angle de 5 degrés. Ensuite la Troposphère et l'Ionosphère sont responsables respectivement du retard de la propagation des signaux et de la réception des signaux. L'Ionosphère est à l'origine de la plupart des erreurs naturelles. Les retards engendrés sont faciles d'être modélisés, alors que les retards engendrés par la Troposphère sont plus difficiles à modéliser car ils dépendent de la température, de la pression et de l'humidité de l'air. D'ailleurs l'erreur peut être issue de la géométrie des satellites visibles. Si les satellites visibles sont très proches dans l'espace, la précision sera moins bonne que si elles sont réparties régulièrement sur une large étendue au-dessus de l'utilisateur. Finalement la synchronisation de l'horloge et du récepteur est importante pour avoir une bonne précision de GPS. Il faut qu'un récepteur soit équipé d'une horloge atomique, identique à celles des satellites [6][7].

3. Démarche suivie

Pour mener à bien le projet, une démarche progressive en quatre temps forts a été adoptée :

- ❑ Une réflexion initiale sur le mode de collecte des données GPS
- ❑ Une analyse des paramètres pouvant influencer la qualité des données GPS par la collecte des données des membres du groupe de PE
- ❑ Une collecte des données GPS d'un panel d'une vingtaine d'élèves de l'école
- ❑ Un travail de conception et de réflexion autour des représentations graphiques et cartographiques possibles pour l'analyse des données

3.1 Choix de l'application pour la collecte des données GPS

La collecte de points de coordonnées longitude, latitude, temps étant au cœur du projet, il est essentiel, au départ, de réfléchir à un mode d'acquisition de données GPS qui puisse convenir.

L'un des objectifs intermédiaires est de réaliser une collecte de données GPS auprès d'un panel d'étudiants. Il faut donc choisir un format d'acquisition dont la prise en main est facile et que l'on peut diffuser au plus grand nombre à moindre coût. D'autre part, on souhaite procéder à une étude sur les erreurs GPS à l'échelle du campus, il faut avoir des précisions temporelles et spatiales suffisantes et une visualisation accessible des données brutes pour conclure rapidement sur les hypothèses faites.

Utiliser une application mobile semble donc être le choix le plus adapté par rapport à des traceurs GPS indépendants. On réalise donc une étude afin de comparer différentes applications de tracking disponibles. Les critères principaux sont les suivants :

- Récupération possible de l'historique des données GPS
- Précision spatiale à l'échelle du campus
- Temps d'échantillonnage des acquisitions à l'échelle de la minute
- Installation possible sous Android et iOS
- Coût faible voire nul
- Visualisation rapide des données brutes depuis l'application

Application	Avantages	Inconvénients
FollowMee	<ul style="list-style-type: none"> Récupération des données au format gpx 	<ul style="list-style-type: none"> Historique des positions payant pour iOS et qui devient payant sous Androïd Précision de la localisation (point à Charrière Blanche au lieu de Comparat) Temps entre chaque acquisition trop long (supérieur à certains temps de trajet à pieds par exemple) Consommation de batterie
Google Maps	<ul style="list-style-type: none"> Données plus précises (incertitude sur la position réduite par rapport à d'autres applications) Possibilité d'ajouter des lieux visités/trajets dans la journée si non relevés par l'application Données collectées par défaut par Google sous Androïd Visualisation de la position en direct sur une carte et a posteriori dans l'historique des trajets Disponible sous Android et iOS et préinstallée sur la majorité des smartphones Facile à prendre en main 	<ul style="list-style-type: none"> Consommation de batterie Envoie de données de localisation à Google (Questions de sécurité, privacy)
GPS Tracker for Android	<ul style="list-style-type: none"> Affichage directement sur une carte 	<ul style="list-style-type: none"> Pas de récupération des données Indisponible sur iPhone Questions de sécurité (développeur inconnu)
Geo Tracker	<ul style="list-style-type: none"> Multitude de format de données (GPX, KML,KMZ) 	<ul style="list-style-type: none"> Consommation de batterie Acquisition des points de données aléatoire (pas de réglage du temps d'acquisition) Questions de sécurité (développeur inconnu)
GPS-Tracker Pro	<ul style="list-style-type: none"> Période d'acquisition réglable Visualisation de la trajectoire sur carte, du mode de transport Notification en cas de faible signal Facile à prendre en main 	<ul style="list-style-type: none"> Consommation de batterie Questions de sécurité (développeur inconnu)

Tableau 1 : Avantages et inconvénients de différentes applications de géolocalisation

Bilan :

A l'issu de ces comparaisons, Google Maps apparaît comme la solution la plus judicieuse. Elle allie facilité, accessibilité et fiabilité (peu de bug), et par rapport à d'autres applications tout aussi performantes, elle est souvent préinstallée et déjà connue de la plupart des personnes. Cette application sera donc retenue pour les collectes de données dans la suite du projet.

3.2 Analyse des paramètres influençant la qualité des données GPS

Le choix de l'application de Tracking réalisé, on effectue une première collecte des données des membres du groupe. Cette étape essentielle a pour but de nous familiariser avec les données à recueillir : format des fichiers (json), marquage temporel, nombre de points obtenus.

En comparant ces données sur différentes périodes et entre les membres du groupe, il apparaît un certain nombre de différences tant en nombre de points qu'en qualité des points acquis. Il semble donc que certains paramètres influencent l'acquisition des données et rendent nos points plus ou moins fiables. Pour déterminer ces paramètres et comprendre leur influence, la démarche suivante a été adopté.

3.2.1 Démarche expérimentale

Afin de comprendre la qualité des données GPS obtenues avec la géolocalisation et l'application Google Maps, on définit :

1. les critères qui définissent la qualité des données
2. la configuration optimale de Google Maps
3. les paramètres dont on étudie l'influence
4. les hypothèses de départ sur les paramètres
5. la démarche suivie pour les expériences

Tout d'abord, on définit les critères qui permettent de dire si les données GPS sont de qualité ou non :

- **le nombre de points** enregistrés sur un chemin fixe - on considère que la qualité augmente avec le nombre de points.
- **la précision géographique** des points - l'écart entre la vraie position et la position mesurée.

La précision temporelle des points - pour une position fixée, l'écart entre le temps réel et le temps obtenu avec le GPS - pourrait aussi être considérée comme un critère de qualité. Cependant d'une part la synchronisation des récepteurs GPS avec les satellites se fait avec la précision d'une horloge atomique, d'autre part, une désynchronisation des horloges du récepteur et de l'émetteur a une influence directe sur la précision spatiale : le calcul de la position spatiale se fait à partir du temps de parcours de l'onde entre l'émetteur et le récepteur : $d = c \times \Delta t$.

3.2.2 Configuration optimale de Google Maps

Avant que l'on ne commence avec les tests sur les hypothèses, on fait connaissance avec l'application choisie : Google Maps. Notamment, on veut maximiser la quantité de points enregistrés par l'application dans chacune de nos expériences.

Les premières mesures réalisées avec Google Maps montrent que l'application enregistre approximativement un point toutes les 15 secondes quand l'application google maps est active et ouverte¹. Par contre, elle enregistre un point approximativement toutes les 30 secondes quand l'application est active mais pas ouverte.

La présence de la connexion internet n'influence pas la quantité de données collectées comme le présente le tableau 1 et la figure 1 qui récapitule les paramètres variés pendant le trajet correspondant.

	Internet	App active	App ouverte
1	non	non	non
2.	non	oui	oui
3.	non	oui	non
4.	oui	oui	non
5.	oui	oui	oui

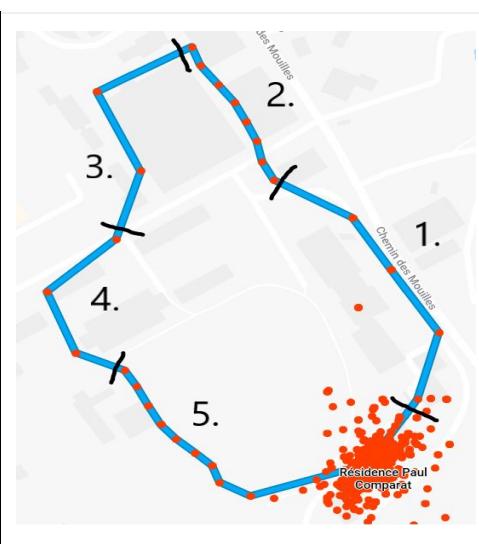


Figure 1 : Trajet obtenu avec Google Maps

Une fois cette information obtenue, on choisit de maintenir l'application Google Maps ouverte pendant toutes les expériences suivantes, afin de maximiser les points collectées. Par ce choix, on s'affranchit de l'influence de Google Maps sur le nombre de points collectés pour la plupart des paramètres présentés dans la partie suivante.

Bilan :

¹ On fait la distinction entre application **active** et application **ouverte** : “l’application est active” veut seulement dire qu’on l’a ouverte, mais qu’on ne va pas forcément la garder sous les yeux pendant le trajet. En revanche, “l’application est ouverte” veut dire qu’on ne laisse pas le portable passer en stand-by pendant le trajet et on garde l’application Google Maps en premier plan.

La configuration optimale de Google Maps pour l'usage de l'application que l'on fait dans notre étude est donc : internet allumé ou non, Maps active et ouverte.

3.2.3 Paramètres considérés et hypothèses de départ

On détermine alors la liste des paramètres pertinents qui peuvent jouer sur ces critères et donc influencer la qualité des données GPS. Les hypothèses qu'on fait sur l'impact de chacun de ces paramètres figurent dans le tableau suivant :

Paramètres	Hypothèses de départ
Paramètres internet	
Si internet : type (wifi/données mobiles)	On a un meilleur résultat avec le Wi-Fi qu'avec les données mobiles.
Si données mobiles : Edge/3G/H/H+/4G	Meilleur est le réseau, meilleures sont les données.
Paramètres téléphone	
Type de téléphone (Androïd/ iPhone)	Pas de différence.
Niveau de batterie	Mieux avec la batterie chargée
Paramètres application	
Beaucoup d'applications ouvertes en parallèle ?	Si on a beaucoup d'applications ouvertes en parallèle, Google Maps ne va pas enregistrer beaucoup de points.
Application de trackage GPS allumée en parallèle ?	Une application de trackage GPS (type OruxMaps ou FollowMee) va forcer l'acquisition de données GPS.
Paramètres extérieurs	
Météo	Pas de différence Sauf si le temps est très mauvais (ex : pluie, orage), nous devrions avoir de plus mauvaises données.
Présence dans un bâtiment	On reçoit moins bien les signaux GPS dans un bâtiment qu'à l'extérieur.
Heure de la journée	Pas de différence
Jour de la semaine	Pas de différence
Nombre de personnes aux alentours	Si le réseau est plus utilisé, on devrait avoir des données GPS de moins bonne qualité.

Tableau 2 : Paramètres influençant la qualité des données acquises et hypothèses sur leur influence

3.2.4 Protocole pour chaque expérience

Pour tester les paramètres établis et valider ou non les hypothèses formulées, on fixe un trajet sur le campus. On définit des points communs à toutes les expériences sur ce trajet pour lesquels l'horaire de passage est mesuré à la main. Ce trajet est appelé Trajet (1).

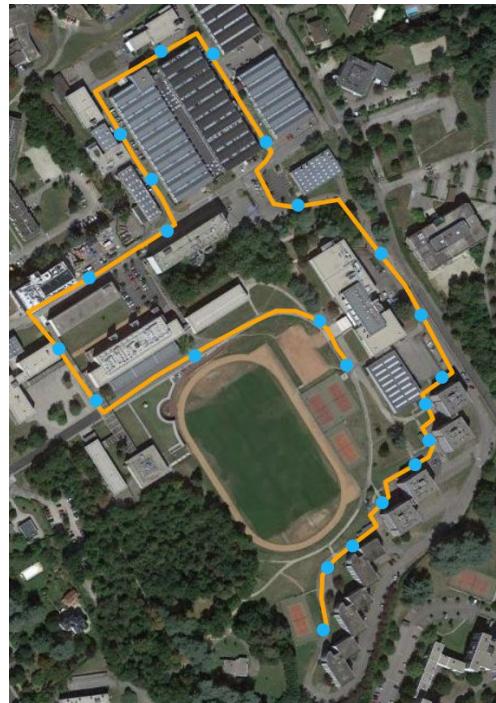


Figure 2 : Trajet (1) en orange avec les points définies en bleu [14]

La démarche suivie pendant chaque expérience est la suivante :

- Au début de chaque trajet on renseigne les états des paramètres dans le tableau. Pour faire un test sur un paramètre, on va faire plusieurs essais en maintenant tous les autres paramètres fixes et en faisant varier seulement le paramètre sur lequel on est en train de tester l'hypothèse.
- Lors du parcours de trajet, on note l'horaire de passage en chaque point (bleu) du Trajet(1), pour pouvoir comparer le parcours avec ce que Google Maps en a retenu.

Les résultats obtenus pour les différents paramètres sont présentes dans la partie 4.1.

Bilan :

Il existe des paramètres qui modifient la qualité des données GPS et donc la précision de l'analyse que l'on pourrait en faire. Il s'agit de tester l'impact propre de chacun de ces paramètres par comparaison. Pour le faire on définit un protocole qu'on suit pendant chaque expérience.

3.3 Démarche informatique

3.3.1 Définition de zones d'intérêt sur le campus

Pour mener une analyse au plus simple de l'ensemble des données GPS récoltées et afin de comprendre les trajectoires sur le campus, on choisit de définir des zones géographiques (carrés, rectangles, etc) sur le campus pour y inclure les points des données. On utiliser Open Street Map qui donne les coordonnées GPS précises, et donc permet de trouver l'équation des droites qui définit des contraintes d'inégalité pour qu'un point appartienne à la zone . En utilisant la méthode “Optimisation Linéaire”, on détermine la zone géographique et on affiche les points utiles dans cette zone. Dans ce cas, le campus a été subdivisé manuellement en sept zones géographiques principales comme le montre la Figure 3.

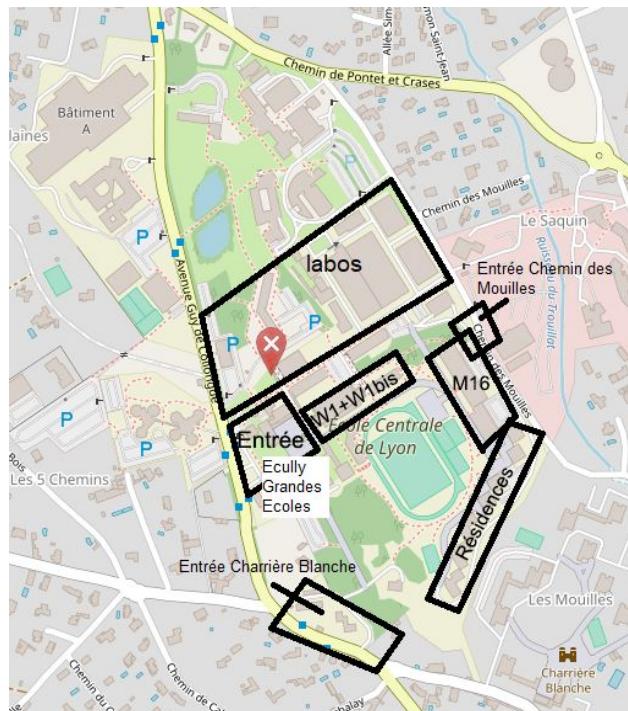


Figure 3 : Division du campus en zones géographiques

3.3.2 Application de la méthode K-means aux données de la collecte

On a fait une hypothèse sur les zones d'intérêt du campus à partir des connaissances pratiques du terrain. Pour valider ou invalider cette hypothèse, on utilise les données récoltées sur le campus pendant la collecte du mois de mars.

La méthode K-means (ou K-moyennes) est une méthode heuristique de partitionnement de données. Elle fait partie des méthodes de "clustering" (en français "regroupement") qui visent à séparer un jeu de données en classes regroupées dans un espace. [1] On applique donc cette méthode aux données récoltées pour associer à chaque point de données une zone de l'espace.

Le principe de la méthode K-means est le suivant :

On a un jeu de points dans un périmètre délimité de l'espace que l'on souhaite partitionner en N zones.

Initialisation :

- On choisit au hasard à l'intérieur du périmètre N points appelés centres.

Itération : Tant que les centres ne sont pas fixes entre 2 itérations :

- On associe à chaque points de donnée le centre le plus proche. Cela forme des zones.
- On recalcule les nouveaux centres qui correspondent aux barycentres des points de chaque zone.

Conclusion :

- On a un partitionnement des données en N zones.

Le programme correspondant est donnée en Annexe. Il applique cette méthode et affiche sur une carte les zones correspondantes pour l'ensemble des données collectées sur le campus de l'Ecole Centrale de Lyon.

En raison du grand nombre de points que l'on considère (1753 points), le principe de base est optimisé pour parvenir à réduire le temps de calcul. En effet, le calcul de distances requiert l'utilisation d'une fonction racine carrée. On compare donc plutôt la distance au carré pour éviter ce problème.

Les modifications apportées sont les suivantes :

- **Conditions initiales** avec la méthode de Forgy [8] : centres initiaux choisis parmi l'ensemble des points de données
- **Calcul des zones** : définition pour chaque centres d'un rayon minimal en pour lequel un point de distance au centre inférieur au rayon minimal appartient forcément à la zone associée au centre. Pour chaque point on calcule d'abord sa distance avec le centre de sa zone précédente, s'il elle est inférieure au rayon minimal, on sait qu'il appartient encore à la zone. Cela minimise la complexité en évitant N-1 calculs de distance (et donc de racine carrée) par points bien placé. Optimisation pertinente dans la mesure où nos points sont agglutinés de manière dense en certains espaces du campus.

Les autres perspectives d'optimisation qui ne sont pas appliquées dans le programme sont les suivantes :

- Choix d'un critère de convergence qui ne soit pas la stricte égalité de centres entre 2 itérations consécutives.
 - Amélioration de la complexité du calcul des nouveaux centres

La méthode dépendant des conditions initiales, on effectue plusieurs test pour différentes conditions initiales et pour différents nombres de zones.

Voici donc dans la suite une illustration de la méthode appliquée aux données brutes de la collecte et affichée sur une carte du campus. Sur la figure 4 l'ensemble des points de la collecte situés dans le périmètre du campus est représenté sur une carte du campus de l'Ecole Centrale de Lyon.

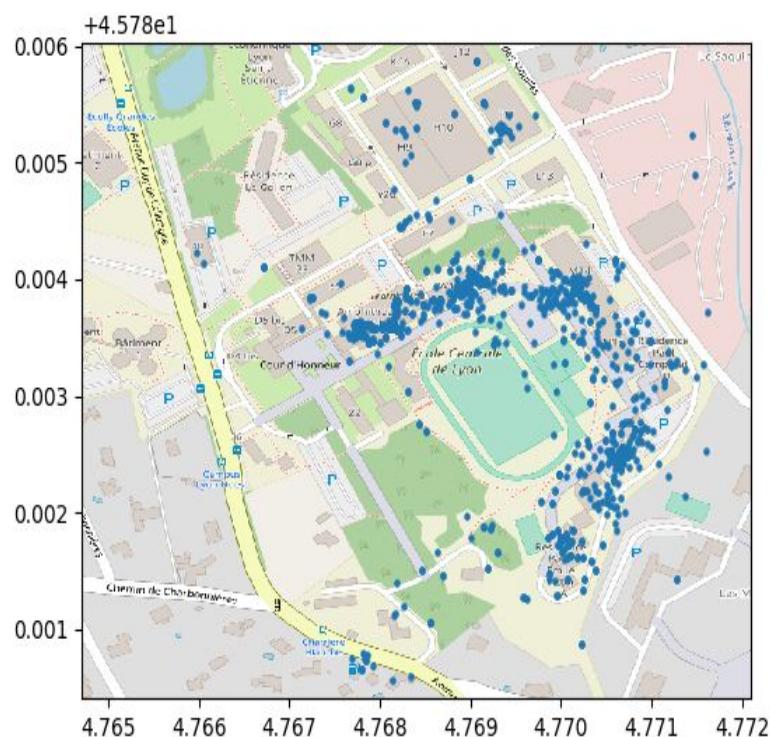


Figure 4 : Ensemble des points collectés auprès des participants

Sur la figure 5, les centres initiaux sont placés de manière aléatoire dans le périmètre des points ainsi leurs zones initiales correspondantes sont représentées (phase d'initialisation).

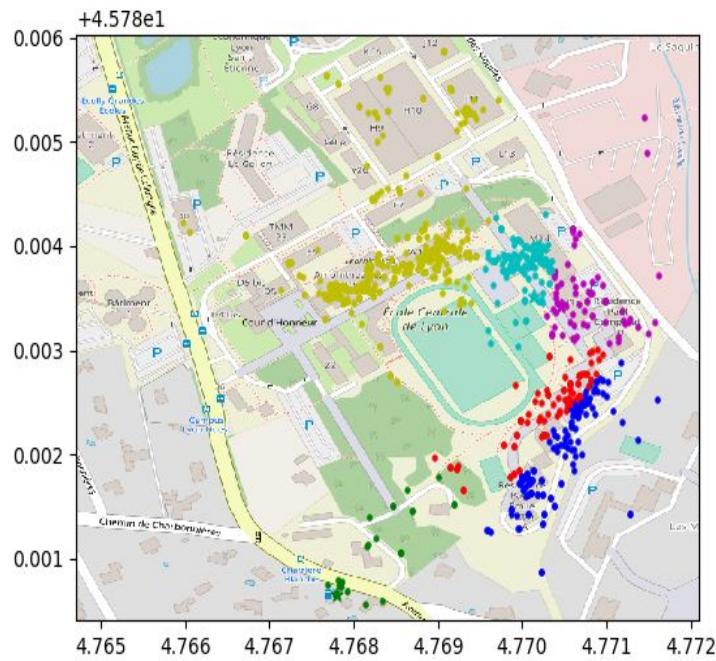


Figure 5 : Initialisation de la méthode des K-means

Les 6 zones finales obtenues pour cette initialisation sont affichées figure 6 :

- Administration, W1-W1bis et Bibliothèque (en jaune)
- Laboratoires (en bleu clair)
- RU, M16 et Gymnase (en violet)
- Résidence Comparat (en rouge)
- Résidence Adoma (en bleu foncé)
- Arrêt de bus Charrière Blanche (en vert)

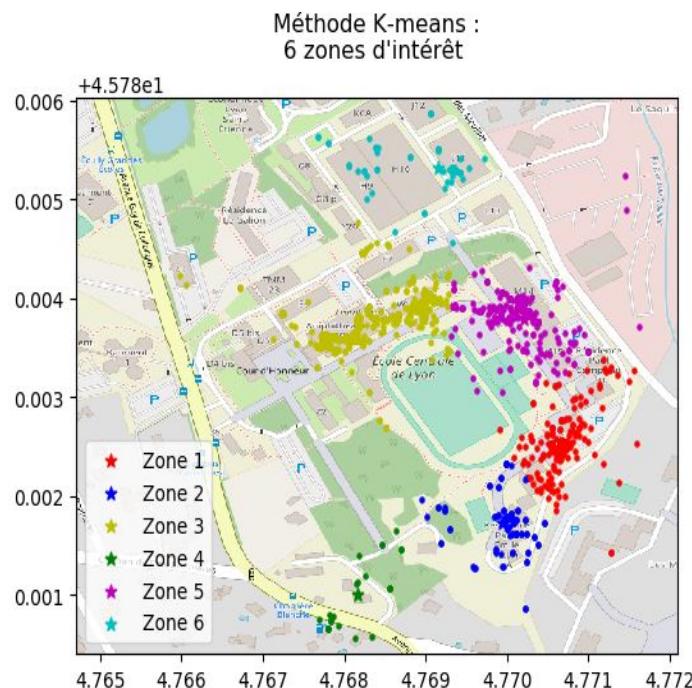


Figure 6 : Résultat final de la méthode pour ces conditions initiales

3.3.3 Visualisation cartographique des données collectées

Une visualisation simple de nos données (figures 20 et 21) nous oblige déjà à remettre en question la qualité de nos données et nous pousse à donc revenir sur notre collecte et à se demander quelles paramètres influencent la collecte de points par Maps.

L'un des objets de notre étude est aussi de commencer à dégager des conclusions des données obtenues. On a accès à des jeux de données GPS comportant les informations de localisation et de temps au format json. On essaye d'induire de nos données des comportements de déplacements. Pour cela nous voulons utiliser des méthodes de représentation de données.

La première option est de visualiser nos données sur des cartes. Nous avons ainsi projeté nos données sur des cartes (figure 20), puis essayé de représenter la densité de ces points sous forme de *heatmap*. Le principe de la *heatmap* ou carte de densité est qu'elle permet d'extraire de la multiplicité des points une répartition de la densité de ceux-ci. La densité de points est représenté sur des cartes par des échelles de couleur et l'affichage des contours de niveau des zones les plus denses.

Pour notre étude, on choisit d'abord de regarder la densité des points accumulés sur toute l'étude à toute heure (figure 22) et sur des plages horaires plus précises (figure 23). Les résultats relatifs à cette visualisation seront présentés dans la partie 4.2.2.

3.3.4 Etablissement d'un graphe de déplacements entre les zones du campus

L'un des enjeux de l'analyse de données de géolocalisation est de pouvoir prédire la destination d'un utilisateur en fonction de ses dernières positions connues. On va ainsi s'inspirer de [9] afin de proposer un modèle de chaîne de Markov simplifié.

Un processus de Markov est un modèle probabiliste qui sert à prédire, parmi plusieurs états possibles, quel est l'état le plus probable après un certain nombre d'itérations de ce processus. Il consiste essentiellement en une matrice des probabilités de transition entre 2 états, ce qui peut aussi être représenté comme un graphe orienté qui admet pour noeuds les états possibles du système et dont les arêtes sont étiquetées par les probabilités de transition entre les états (c'est cette représentation qu'on choisira dans ce rapport). Pour nous, les états seront bien sûr des positions.

Pour ce faire, nous allons faire quelques hypothèses:

- On supposera que la position de chaque individu est unique pour chaque heure
- On approxime la position des individus par des zones prédéfinies "typiques" sur le campus
- On choisit la zone dans laquelle se situe l'individu à une heure donnée en choisissant la zone dans laquelle on trouve le plus de points à cette heure-là.

Ici on peut remarquer que nous différons déjà de [1] car nous n'employons pas d'algorithme de clustering afin de définir les zones, même si l'on aurait pu le faire comme on l'a vu en 3.3.2. De plus, on "élague" nos données en réduisant à une position pour chaque heure pour chaque participant. Cela nous permet de limiter l'influence des utilisateurs qui nous ont donné le plus de points.

En ayant ainsi déterminé la position de chaque individu à chaque heure, on peut maintenant déterminer à chaque heure h vers quel zone l'individu se déplace pour l'heure $h + 1$, on a la liste des transitions partant de chaque zone à l'heure h et la destination de ces transitions.

On peut alors déterminer, pour chaque jour, une probabilité de passage $p_{i \rightarrow j}^h$ de $zone_i$ à $zone_j$ en prenant :

$$p_{i \rightarrow j}^h = \frac{1}{n_i} \sum_{\text{individu } k} \chi(\{k \text{ est allé de } zone_i \text{ à } zone_j \text{ entre } h \text{ et } h + 1\}).$$

où χ est l'indicatrice et les n_i sont le nombre de transition partant de $zone_i$.

Cependant, notre jeu de données étant peu étendu, nous préférons essayer d'accumuler nos données sur les cinq jours de notre étude, ainsi $p_{i \rightarrow j}^h$ peut être interprétée comme une probabilité typique pour qu'un individu, se trouvant dans la $zone_i$ à l'heure h , qu'il se trouve dans la $zone_j$ à l'heure $h + 1$, et ce n'importe quel jour de la semaine.

On obtient alors un graphe qui représente cette chaîne de Markov (figure 24).

On a ainsi décrit une démarche complète qui permet d'obtenir un modèle de prédiction : on dégage d'abord des zones d'intérêt (par des connaissances préalables sur les données ou en les déduisant des données par des algorithmes de clustering) puis on construit notre modèle de chaîne de Markov en fonction du nombre de positions précédentes que l'on prend en compte.

Bilan :

Le campus a été découpé en zones géographiques principales dans le but de caractériser les trajectoires sur celui-ci, et ce, notamment, grâce à la fréquentation des dites zones. On voit dans la partie suivante que nos résultats correspondent plutôt bien à nos attentes.

4. Résultats obtenus

4.1 Résultats de l'analyse des paramètres influençant la qualité de la collecte

La première hypothèse testée est l'indépendance des jours et des créneaux horaires. La raison étant que si l'indépendance est vraie les autres expériences peuvent être faites tous les jours à tous les créneaux horaires car ce ne sont plus des paramètres pertinents.

4.1.1 Influence des paramètres extérieurs

4.1.1.1 Indépendance des jours de la semaine

Le Trajet (1) est parcouru avec trois téléphones différents chaque soir pendant une semaine. Tous les paramètres sauf le jour de la semaine ont été maintenus fixes, comme listé dans le tableau présenté en Annexe p.49. Après chaque expérience les données obtenues avec Google Maps sont comparées avec le vrai trajet.

La précision, pour un téléphone fixe, ne varie pas significativement avec les jours de la semaine comme on peut le voir sur les figures 7 à 9 où le tracé bleu correspond à la trace GPS du trajet (1) enregistrée par Google Maps.



Figures 7 à 9 : de gauche à droite : mesures obtenues avec les mêmes paramètres le mercredi 16/05/2018 et le vendredi 18/05/2018 entre 23h et minuit et comparées avec le Trajet(1).

4.1.1.2 Indépendance des créneaux horaires

La même démarche est appliquée pour des créneaux horaires différents. On constate que l'heure de la journée n'est pas non plus un paramètre pertinent. Comme représenté figure 10, on ne constate pas de divergence majeure entre les deux traces GPS.

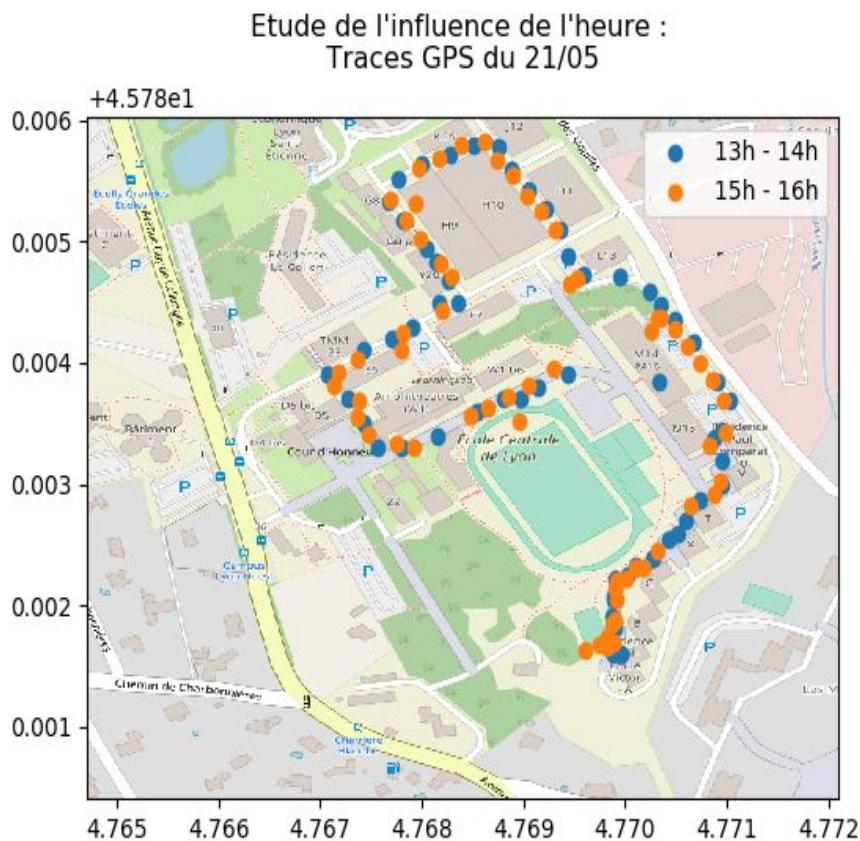


Figure 10 : exemple de deux trajets faits le lundi 21/05/2018, le premier entre 13h40 - 13h54 et le deuxième entre 15h44 - 16h00

Bilan :

Les hypothèses suivantes sont confirmées :

- le jour de la semaine n'a pas vraiment d'influence sur la précision des données.
- le créneau horaire n'a pas vraiment d'influence sur la précision des données.

Ces deux paramètres ne sont donc plus pris en compte dans les expériences suivantes.

4.1.1.3 Influence de la météo

La théorie dit que les signaux GPS deviennent moins fiables en présence d'un ciel trop nuageux. [10]

Les couches d'atmosphère réfractent les signaux GPS de diverses manières. Bien que le retard atmosphérique dans l'équilibre hydrostatique peut être modélisé adéquatement, l'aspect délicat est la partie causée par la vapeur d'eau dans les parties inférieures des troposphères. Notamment, la vapeur d'eau varie de façon imprévisible, et souvent pas de la même manière qu'on peut mesurer sur la surface avec un lecteur d'humidité. Par conséquent, il est impossible de modéliser correctement le retard.

On réalise plusieurs essais sous la pluie, et avec un ciel nuageux pour confirmer cette hypothèse. Les résultats obtenus divergent. Des expériences montrent le même résultat que ceux avec le ciel dégagé. En revanche, on fait des expériences où aucun point n'est relevé.

Bilan :

Bien qu'on obtienne des résultats divergents, on décide de confirmer l'hypothèse donnée par la théorie - la météo a une influence non négligeable sur la précision des données collectées.

4.1.2 Influences des paramètres liés à l'application Google Maps

4.1.2.1 Influence du nombre des autres applications actives

Avec un nombre important d'applications en parallèle, la précision de Google Maps est toujours la même - si l'application est en premier plan. Sur la figure ci-dessous on note que pendant la partie du trajet derrière et devant M16 le nombre de points relevés avec les application actives est inférieur à celui qu'on relève dans le cas contraire. Ceci est la conséquence du fait que pendant ces parties, l'application Google Maps était active mais pas ouverte.

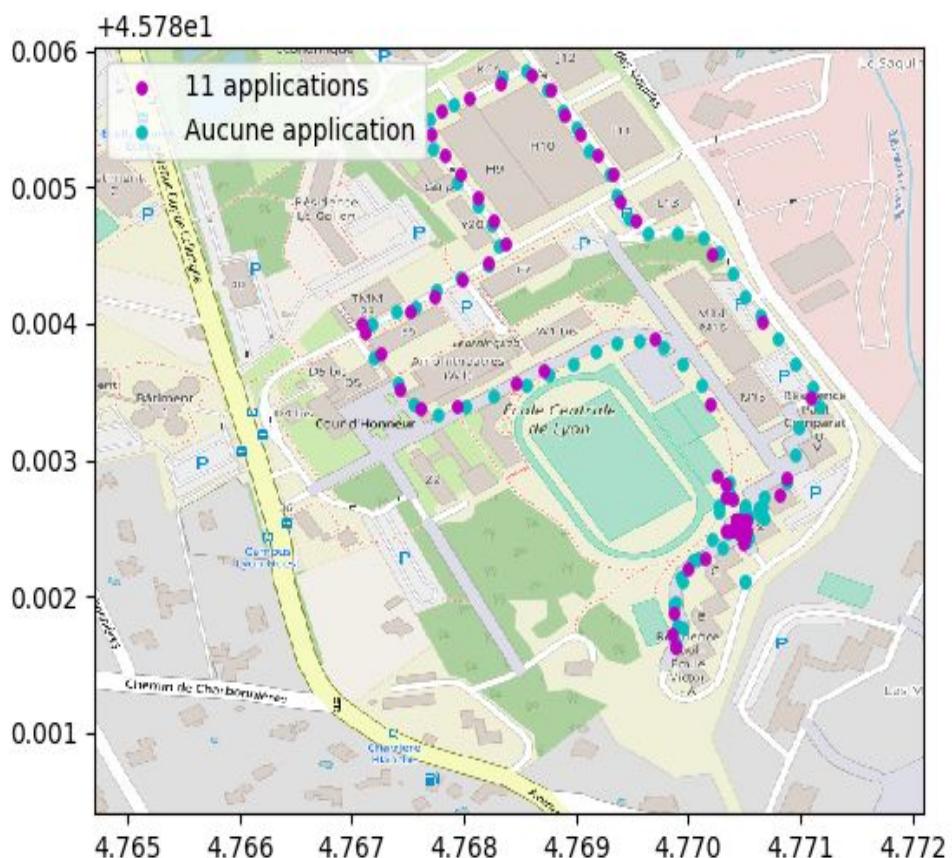


Figure 11 : Tracé GPS du 16/05 et du 07/06 pour étude de l'influence des applications ouvertes en parallèle

4.1.2.2 Influence des autres applications de trackage

Pour voir une différence dans la quantité de points relevés, pendant ce trajet on a laissé Google Maps actif mais pas ouvert.

On note une nette amélioration du relevé des points lorsqu'une application de traçage autre que Google Maps est active en parallèle (et seulement cette application, en l'occurrence FollowMee). Davantage de points sont relevés et coïncident avec l'heure réelle de passage sur le lieu. L'esquisse du trajet proposé par Google Maps devient alors plus vérifique et même meilleure que pour le trajet de référence sans aucune application en parallèle, comme on peut voir sur la figure ci-dessous.

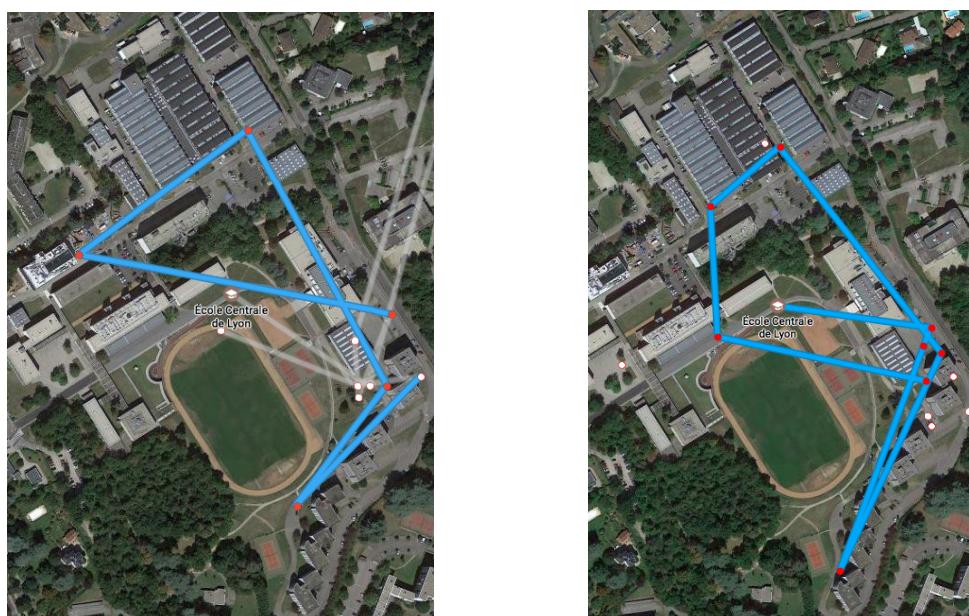


Figure 12 et 13: à gauche le trajet habituel, à droite le trajet réalisé avec l'application FollowMee ouverte

Bilan :

Si l'application Google Maps est ouverte en premier plan, les applications ouvertes en parallèle ne jouent pas un rôle crucial. Au contraire, si une seule autre application est ouverte en parallèle et que celle-ci fait également du relevé de coordonnées GPS en temps réel, le trajet est mieux relevé par Google Maps.

4.1.3 Rôle des paramètres internet

Malgré la popularité de la localisation basée sur le GPS, ses performances dans les environnements intérieurs sont limitées. Cela est dû à la mauvaise pénétration des signaux GPS à l'intérieur des bâtiments et à l'absence fréquente de systèmes de localisation intérieure. C'est pour cela que la géolocalisation fonctionne mal voire très mal dans les bâtiments, notamment quand on est loin des fenêtres, ou dans l'ascenseur.

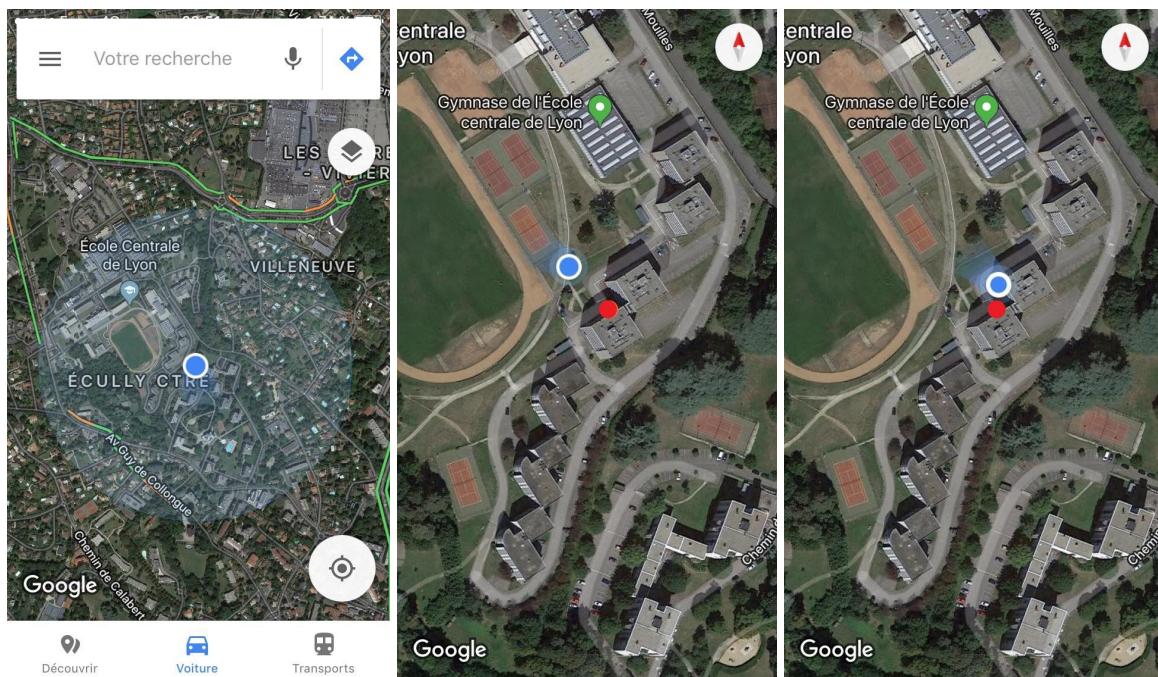


Figure 14 à 16 :

- (1) Le cercle d'incertitude² qui apparaît quand on se trouve dans le couloir du bâtiment X,
- (2) l'incertitude disparaît instantanément dès qu'on s'approche d'une fenêtre, mais le positionnement n'est pas tout à fait précis - en rouge la position exacte.
- (3) Après un temps de transition d'environ 5 secondes on peut voir que le positionnement obtenu avec le GPS s'approche de la position exacte.

² On définit la précision horizontale comme le rayon de confiance de 68%. En d'autres termes, si on dessine un cercle centré à la latitude et la longitude de cet emplacement, et avec un rayon égal à la précision, alors il y a une probabilité de 68% que l'emplacement vrai est à l'intérieur du cercle. [11][12]

On se demande donc : est-ce qu'avec la connection internet on arrive à se géolocaliser mieux dans les bâtiments³ ?

Pour faire des essais avec le WiFi comme avec la 4G, on choisit de faire des trajets dans les bâtiments du campus où l'on a accès au réseau WiFi de l'école - les bâtiments W1 et W1bis. Le réseau WiFi est aussi fort dans les bâtiments H9 et H10, mais nos expériences montrent que la géolocalisation dans ces bâtiments est déjà très précise sans internet, car comme dans le gymnase le toit est percé des nombreuses fenêtres.

Les essais faits avec les connections WiFi et 4G montrent qu'il n'y a pas d'amélioration. Le GPS n'arrive toujours pas à nous capter dans les bâtiments, et il se met à jour une fois sorti du bâtiment.



Figure 17 et 18 : pendant un essai où on traverse le W1bis et W1 en passant de l'entrée est du W1bis : une fois entrés dans W1 bis le GPS ne nous voit pas, le point bleu reste dehors le bâtiment, et il se remet à jour seulement une fois sortis de W1bis, à l'entrée du W1.

³ Dans les dernières années plusieurs systèmes commerciaux de positionnement intérieur (IPS - indoor positioning system) sont apparus sur le marché. Ces systèmes permettent de localiser des objets ou des personnes à l'intérieur d'un bâtiment en utilisant des ondes radio, des champs magnétiques, des signaux acoustiques ou d'autres informations sensorielles collectées par des appareils mobiles. Il n'y a pas encore une norme pour un système IPS. Google aussi a développé sa version [13].

Dans un deuxième essai on se promène dans le W1, en passant par différents étages. Les résultats donnent la figure ci-dessous.

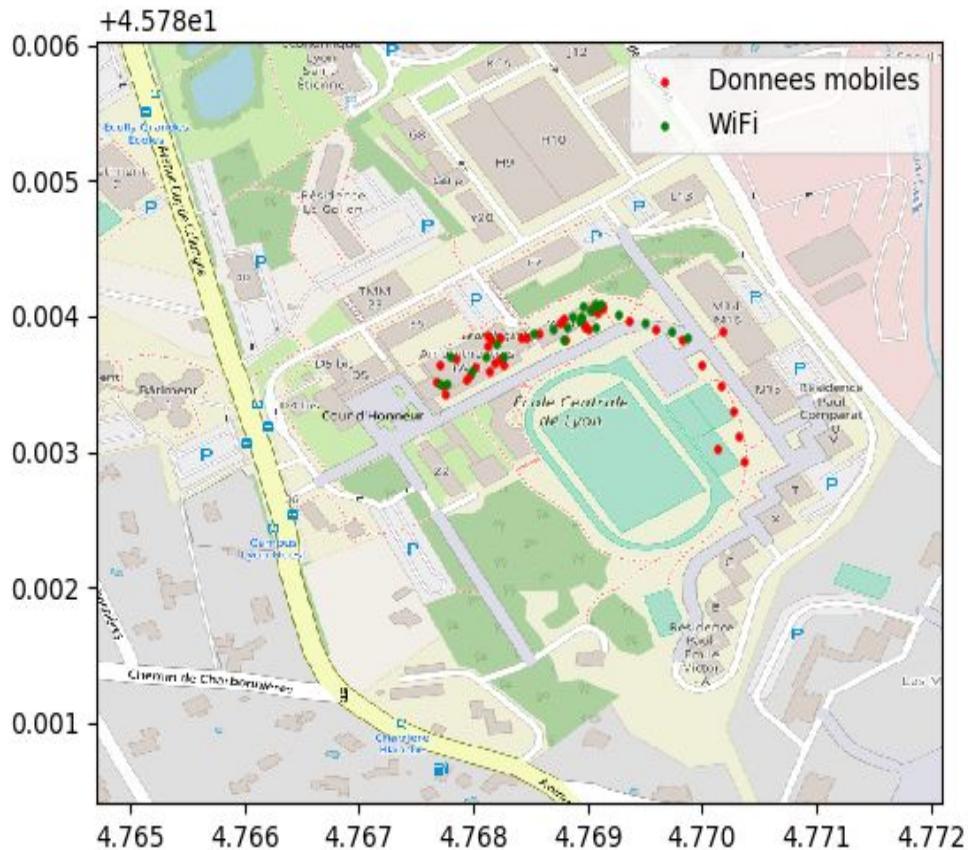


Figure 19 : Tracé GPS du 06/06 entre 16h et 18h pour étude du rôle d'internet

Le fait qu'on est capable de relever des points dans le bâtiment se justifie avec la présence dénormes fenêtres. Les données obtenues avec le 4G et le WiFi ne diffèrent pas suffisamment pour y traiter des conclusions.

Bilan :

On ne remarque pas d'augmentation de précision lorsqu'on allume internet dans le cas de mesures à l'extérieur. Les mesures réalisées à l'intérieur des bâtiments sont aussi imprécises avec que sans la connection internet. Malgré la présence de différents logiciels pour se tracker à l'intérieur, on ne les étudie pas, car on n'utilise pas ces applications pour collecter les données dans le projet.

4.1.4 Influence des personnes alentours

On a mené plusieurs essais pendant des HH sur le campus. On ne remarque pas de comportements aberrants. Il ne faut pas oublier qu'on n'a pas les informations sur le nombre de personnes présents, ni sur les utilisateurs de WiFi/4G/Google Maps actives, par conséquent on ne peut pas vraiment tirer de conclusion.

4.1.5 Influence du niveau de batterie

On a remarqué qu'avec une batterie plus faible de 2% le portable arrête de se tracker. Sinon, la précision du GPS est indépendant du niveau de batterie.

4.1.6 Influence du téléphone en lui-même : modèle, Android/ iOS, version

La carte GPS présente dans tous les portables utilisés dans les expériences est la carte A-GPS ou Assisted GPS card. La partie "assistée" de A-GPS signifie qu'au lieu de laisser le chipset GPS pour comprendre son emplacement sur son propre, une autre partie de l'appareil donne le chipset GPS un indice où commencer à regarder⁴. Ainsi A-GPS obtient la position plus rapidement que le GPS traditionnel, et avec moins de batterie.

Android et iOS à la fois viennent donc avec la carte A-GPS et avec Google Maps. Par contre, les caractéristiques incluses dans Android quand il s'agit de cet application sont beaucoup plus avancés. Cela pourrait être l'explication pourquoi, à parité des paramètres, on a systématiquement plus de point relevés sur les portables avec l'interface Android que iOS.

⁴ Ceci vient typiquement par l'intermédiaire de savoir l'emplacement des stations de WiFi voisines ou des tours cellulaires voisines.

Résumé des résultats obtenus

Paramètre	Hypothèse de départ	Conclusion
Paramètres internet		
Si internet : type (wifi/données mobiles)	On a un meilleur résultat avec la Wi-Fi qu'avec les données mobiles.	Faux. Dans notre étude, on n'a pas remarqué une amélioration des données lorsqu'on réalisait la collecte avec internet allumé.
Si données mobiles : Edge/3G/H/H+/4G	Meilleur est le réseau, meilleures sont les données.	Faux. Dans notre étude, on n'a pas remarqué une amélioration avec internet, l'étude des différentes réseaux perd son intérêt.
Paramètres téléphone		
Type de téléphone (Androïd/ iPhone)	Pas de différence.	Faux. Dans nos expériences les portables avec interface Android ont relevé systématiquement plus de points que ceux avec l'interface iOS.
Niveau de batterie	Mieux avec la batterie chargée	Faux. Le niveau de batterie ne montre pas l'influence sur la qualité de données collectées. Mais, avec une batterie extrêmement faible (<2%) le téléphone arrête de relever les points.
Paramètres application		
Beaucoup d'applications ouvertes en parallèle ?	Si on a beaucoup d'applications ouvertes en parallèle, Google Maps ne va pas enregistrer beaucoup de points.	Faux. S'il est en premier plan, Google Maps se montre aussi performant avec un nombre important d'applications actives.
Application de trackage GPS allumée en parallèle ?	Une application de trackage GPS (type OruxMaps ou FollowMee) va forcer l'acquisition de données GPS.	Vrai. La quantité de points relevés augmente lorsqu'une autre application de trackage est ouverte en parallèle.

Paramètres extérieurs		
Météo	Pas de différence Sauf si le temps est très mauvais (ex : pluie, orage), nous devrions avoir de plus mauvaises données.	Vrai. C'est que avec un temps très mauvais qu'on va remarquer la différence entre la qualité des données relevées. la vapeur d'eau varie de façon imprévisible et cause des retards pas modélisables des signaux GPS.
Présence dans un bâtiment	On capte moins bien la géolocalisation dans un bâtiment qu'à l'extérieur.	Vrai. Le GPS ne marche pas à l'intérieur. Cela est dû à la mauvaise pénétration des signaux GPS à l'intérieur des bâtiments
Heure de la journée	Pas de différence	Vrai.
Jour de la semaine	Pas de différence	Vrai.
Nombre de personnes aux alentours	Si le réseau est plus utilisé, on devrait avoir des données GPS de moins bonne qualité.	On ne peut pas dire. Dans notre étude, on n'a pas remarqué une différence, mais peut-être que le nombre de personnes aux alentours n'était pas suffisamment élevé.

Tableau 3 : Résumé des résultats obtenus avec l'analyse des paramètres

4.2 Présentation des différentes visualisation obtenues

4.2.1 Zones d'intérêt issues de la méthode K-means

Résultats

Comme les résultats finaux de la méthode dépendent des conditions initiales, on a obtenus plusieurs types de partitionnements.

Nombre de zones	Combinaison de zones	Affichage
5	<ul style="list-style-type: none"> - Administration, W1 et bibliothèque, Charrière Blanche - W1bis et laboratoires H9, H10, F7 - Laboratoires I11, J12 - RU, M16, Gymnase, X, T - Résidences sauf X et T 	<p>Méthode K-means : 5 zones d'intérêt</p>
6	<ul style="list-style-type: none"> - Laboratoires - Administration, W1, W1bis, Bibliothèque - RU, M16, Gymnase - Résidence Comparat - Résidence Adoma - Arrêt de bus Charrière Blanche 	<p>Méthode K-means : 6 zones d'intérêt</p>

7	<ul style="list-style-type: none"> - Administration - W1, F7 - W1bis - Laboratoires - RU, M16, Gymnase - Résidence Comparat - Résidence Adoma et Charrière Blanche 	<p style="text-align: center;">Méthode K-means : 7 zones d'intérêt</p> <p style="text-align: center;"> ★ Zone 1 ★ Zone 2 ★ Zone 3 ★ Zone 4 ★ Zone 5 ★ Zone 6 ★ Zone 7 </p>
---	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Tableau 4: Les différents partitionnements obtenus avec la méthode k-means sur l'ensemble des données collectées

En fonction du nombre de zones fixé, on obtient des combinaisons de zones différentes. On remarque cependant dans ces résultats et pour d'autres conditions initiales que certaines zones d'intérêt ressortent de manière plus récurrente : laboratoires, W1, W1bis, résidence Comparat, résidence Adoma, Charrière Blanche, RU-M16-Gymnase.

Les résultats obtenus avec la méthode K-means ont donc tendance à confirmer l'hypothèse faite sur les zones d'intérêt est plutôt confirmer. Il est donc possible de choisir les zones d'intérêt pour l'étude comme une combinaison de ces résultat

Bilan :

La méthode K-means tend à confirmer l'hypothèse faite sur les zones d'intérêt du campus. Les zones d'intérêt retenues dans la suite de l'étude sont donc : W1-W1bis, Laboratoires, RU-M16-Gymnase (bâtiment M), Résidence, Charrière Blanche et entrée Campus Lyon Ouest. Cette dernière est ajoutée pour permettre d'étudier les entrées-sorties des étudiants sur le campus.

4.2.2 Représentations cartographiques

Avant d'effectuer des analyses sur nos données, nous souhaitions visualiser celles-ci de la façon la plus simple possible.

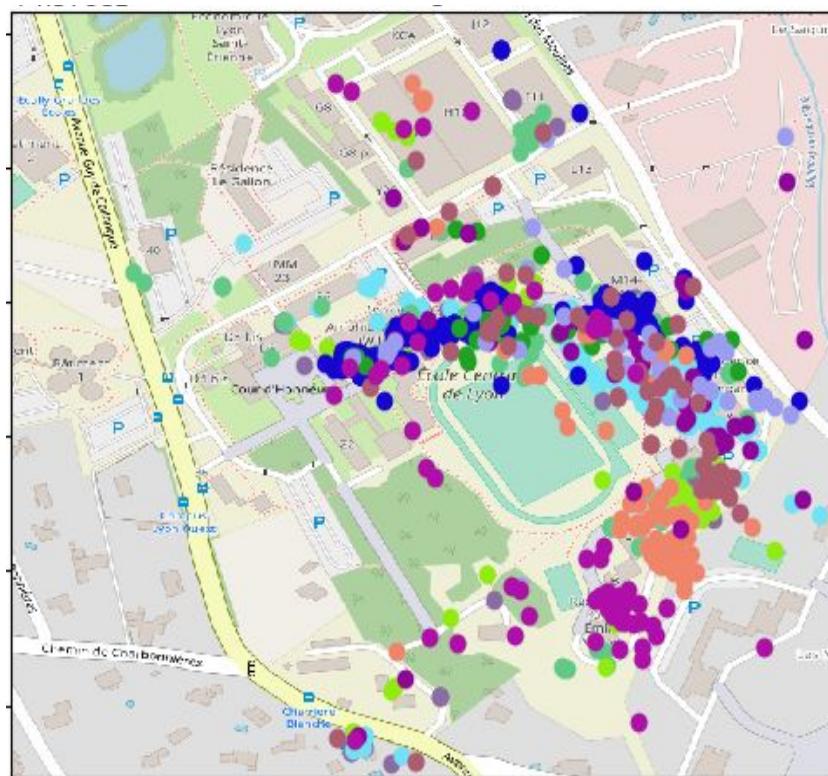


Figure 20 : Ensemble des points récupérés pendant la collecte, une couleur par participant

Si on ne distingue pas de tendance ici, on peut au moins remarquer que les points se répartissent essentiellement en croissant autour du terrain central au niveau des bâtiments de cours, les bâtiments M, les résidences et les laboratoires.

Si l'on regarde plus spécifiquement les zones "interfaces" avec l'extérieur du campus, on voit notamment que c'est le bus à Charrière Blanche qui semble être l'interface privilégiée.

Cependant, on peut aussi se demander à quoi ressemble la qualité des données pour les différents individus de l'étude :

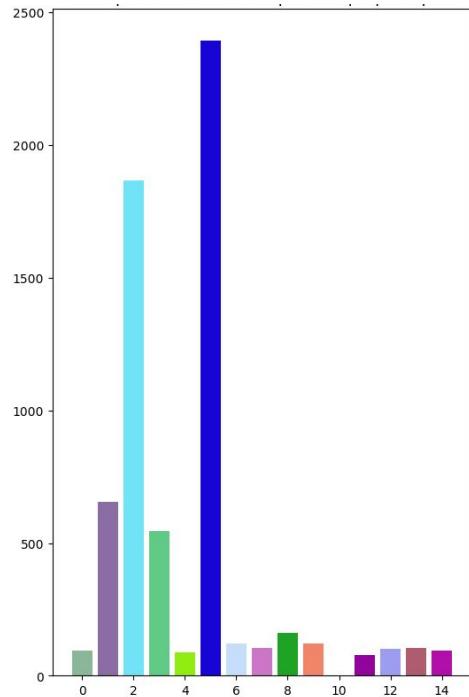


Figure 21 : Nombre de points par participants

On voit alors rapidement un problème avec l'étude : le nombre de points récoltés par chaque participant est très inégal. C'est ce constat qui nous amène à vouloir comprendre plus précisément les paramètres influant sur la qualité des données GPS récoltées (partie 4.1).

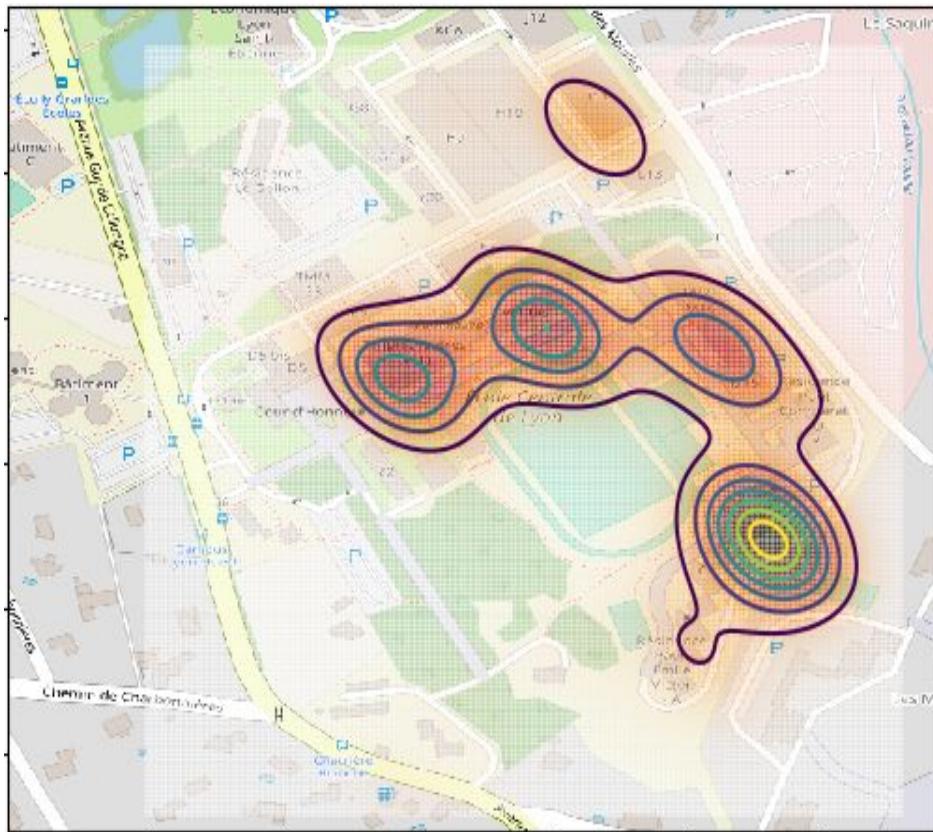


Figure 22 : Heatmap des données de tous les jours pour toutes les heures

Cette première représentation des densités de points nous permet de mettre en lumière 5 pôles principaux sur le campus :

- bâtiment W1
- bâtiment W1bis
- bâtiment M
- Résidences (Adoma et Comparat)
- labos (particulièrement I11)

Ces zones correspondent en grande partie aux zones définies dans la partie 3.3.1, ce qui nous fournit déjà une validation des zones que l'on pouvait prévoir comme étant les plus fréquentées.

On décompose en 6 tranches de deux heures, de 8h à 20h :

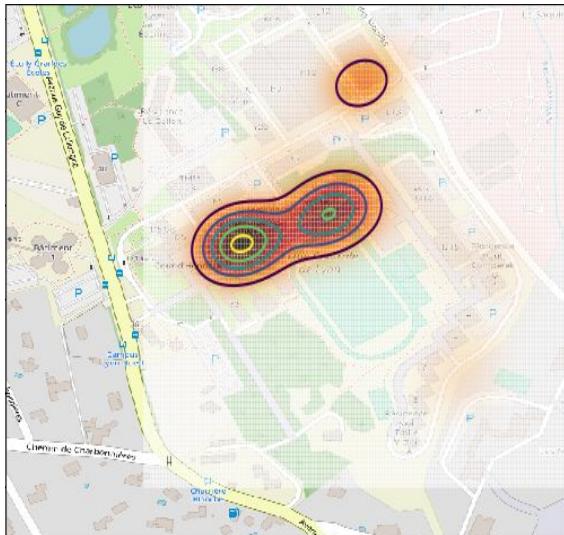


Figure 23.a: Entre 8h et 10h

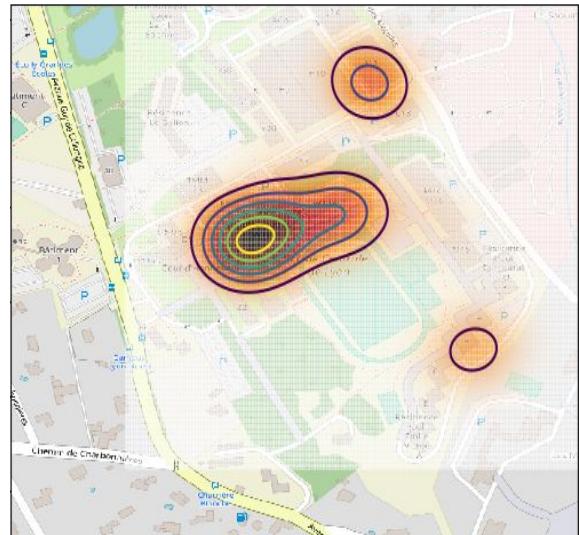


Figure 23.b: Entre 10h et 12h

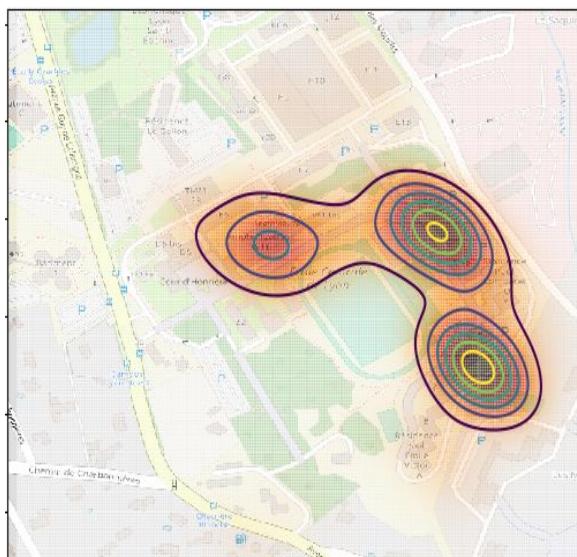


Figure 23.c: Entre 12h et 14h

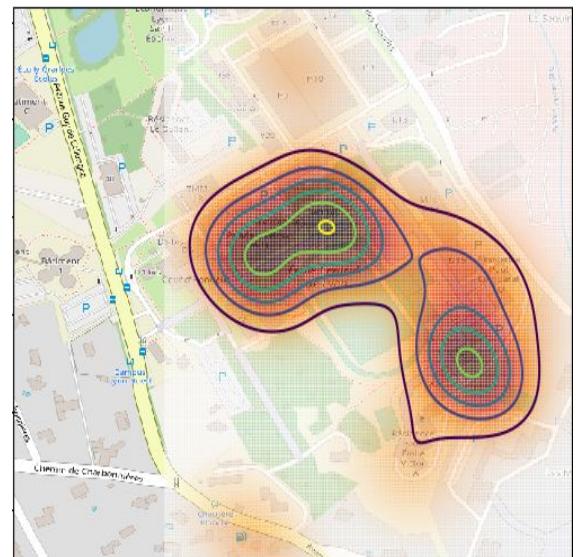


Figure 23.d: Entre 14h et 16h

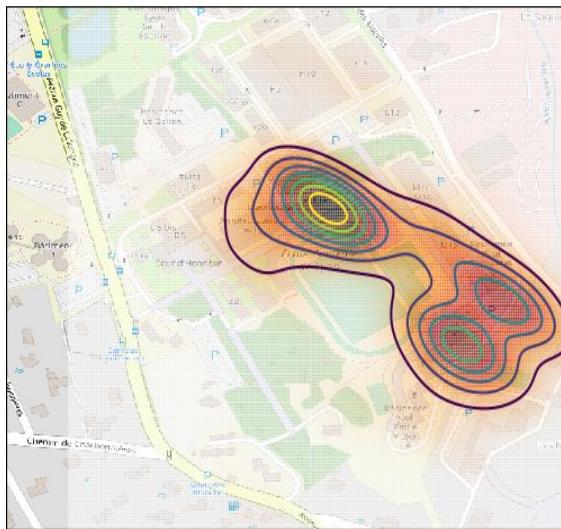


Figure 23.e: Entre 16h et 18h

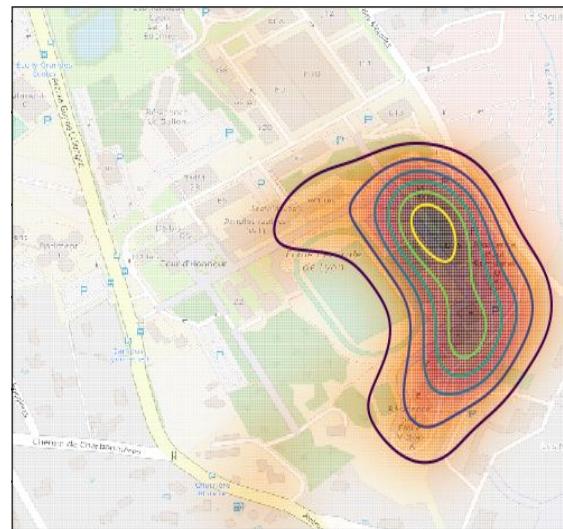


Figure 23.f: Entre 18h et 20h

Figure 23 : Heatmaps montrant la densité, le jeu de points ayant été découpé par tranches de deux heures, de 8h à 20h, dans l'ordre de gauche à droite et de haut en bas

On peut voir très distinctement un inversement de la répartition de densité au cours de la journée : si durant la matinée les points se concentrent exclusivement sur les bâtiments W1 (cours et TDs) et de labos (TPs), on voit que sur le midi les points sont davantage répartis entre le W1, le bâtiment M et les résidences. Entre 14h et 18h, ils se répartissent entre les bâtiments de cours et les résidences, puis dans la soirée du côté des résidences et du bâtiment M, zones de vie étudiante et associative.

4.2.3 Graphe de déplacements entre les zones d'intérêt

Voici les représentations des graphes de transition entre les zones prédéfinies, comme décrits dans la partie 3.3.4 :

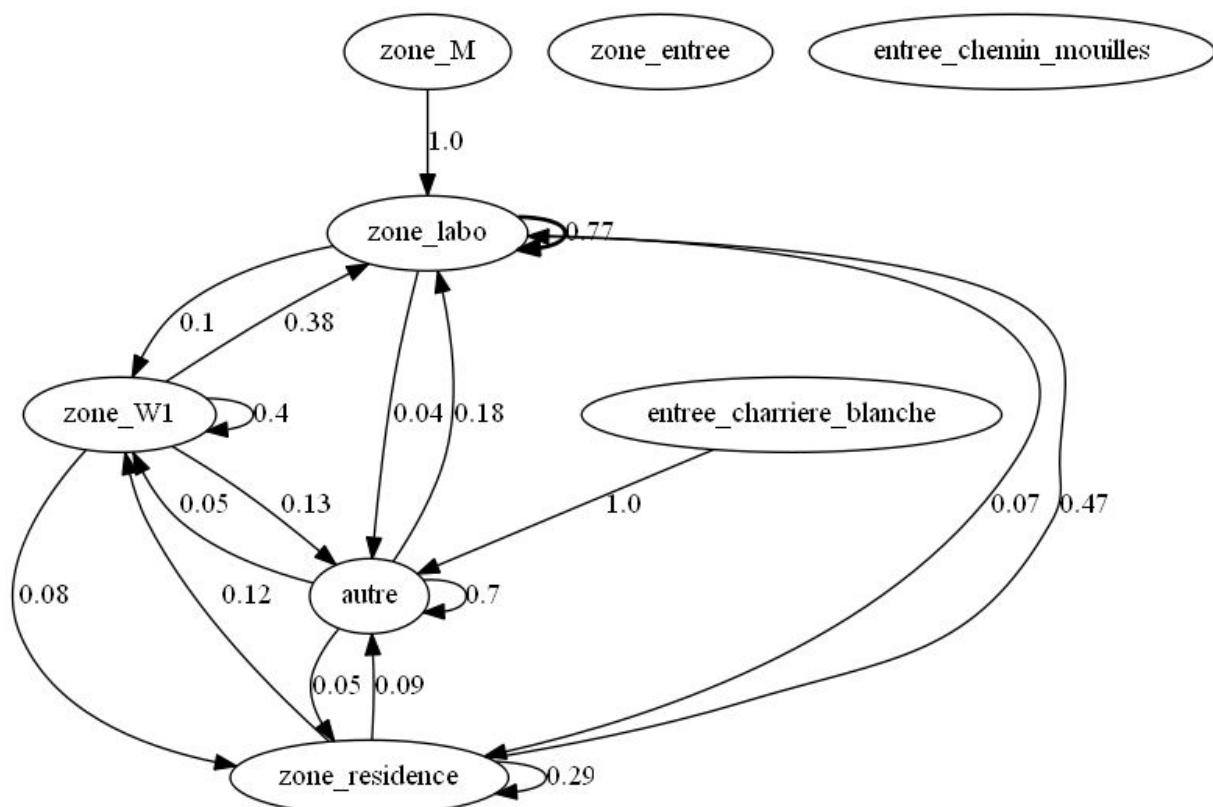


Figure 24 : Graphe de transition pour les données accumulées sur la journée

La figure montre le graphe obtenu via le procédé décrit dans la partie 3.3.4 pour les données accumulées sur la journée. Il est à noter que par soucis de lisibilité, les arêtes étiquetées par 0 ont été effacées. Les zones sont marquées de façon compréhensible par leur dénomination, la zone “autre” ayant été définie comme tout l'espace restant.

Que peut-on remarquer sur ce premier graphique ? Pour les zones les plus étendues (bâtiments de cours, résidences, labos) un individu a une forte chance de rester dans la zone où il était. Cela peut s'expliquer par le fait que les cours à l'ECL sont organisés par créneaux de deux heures. On peut aussi noter qu'aucun individu n'a été assigné aux zones *zone_entree* et *entree_chemin_mouilles*, ce que l'on pouvait prévoir au vue de notre façon de définir les transitions (partie 3.3.4) : on s'intéresse aux transitions entre l'heure h et l'heure $h+1$, or les entrées sont des zones de transition vers ou hors du campus et les étudiants n'y restent pas longtemps.

On peut également limiter notre étude à certaines heures :

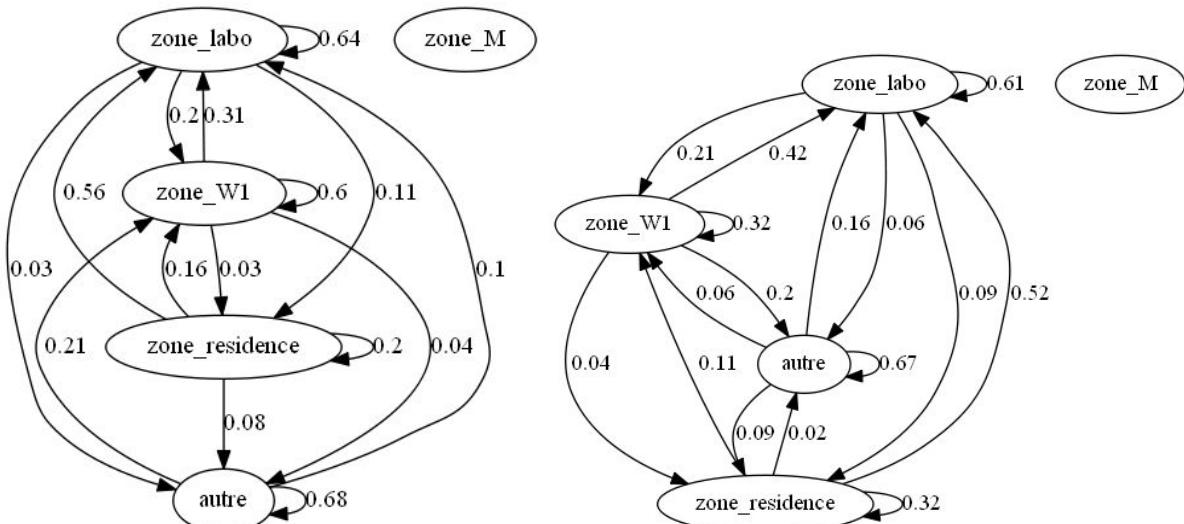


Figure 25 : Graphes des transitions de 7h à 12h et de 13h à 17h

Les principales différences entre la matinée et l'après-midi sont que :

- les individus ont légèrement moins tendance à rester dans les résidences la matinée
- les individus ont moins tendance à rester au W1 durant l'après-midi

On a ainsi obtenu un modèle qui paraît correspondre bien à l'idée préconçue que l'on avait sur le campus. Il peut cependant être affiné, notamment en distinguant une catégorie "autre" d'une catégorie "autre sur le campus Lyon Ouest". Il faudrait aussi essayer d'implémenter un modèle similaire, prenant cette fois en compte les 2 ou 3 dernières positions connues, afin de pouvoir prédire avec plus de précision une prochaine position.

Conclusion

Le projet a pour ambition de comprendre les paramètres influençant la précision du traçage GPS à l'échelle du campus et de visualiser les flux et les densités d'étudiants sur le campus par des méthodes numériques, le tout en s'appuyant sur la collecte de données de géolocalisation.

Les expériences menées permettent de déterminer certains paramètres influençant la précision du traçage GPS par l'application Google Maps sur un téléphone portable. Ainsi il apparaît qu'une météo mauvaise, l'utilisation de certains modèles de téléphones et le fait d'être à l'intérieur de certains bâtiments soient des éléments perturbateurs de la bonne réception des signaux GPS. De plus la fréquence de leur enregistrement peut être améliorer par l'utilisation en parallèle d'une autre application de traçage GPS.

Le travail numérique donne plusieurs méthodes d'analyse des données et de visualisations de celles-ci. La méthode K-means permet de calculer les zones d'intérêt du campus à partir des points de coordonnées récoltés pendant la collecte auprès des étudiants. Les heatmaps mettent en évidence les densités de personnes en fonction des créneaux horaires faisant ainsi émerger des zones d'affluence. Les graphes de transition entre zones d'intérêt montrent les probabilités de déplacement des étudiants entre ces zones.

En somme, les différents axes du projet donnent des éléments qualitatifs à propos de la géolocalisation à une petite échelle géographique et des méthodes algorithmiques servant à visualiser des flux et densités de populations. Cette étude a vocation à être étendue à un effectif plus large d'étudiants pour valider les premiers résultats obtenus. Les nouveaux résultats peuvent être utilisés dans divers contextes comme un réaménagement des lieux de vie du campus ou encore à l'appui d'une demande d'ajustement du réseau de transports publics. Cependant, par sa collecte sur un effectif réduit, le projet s'affranchit des enjeux liés à l'anonymisation des données mais aussi des contraintes de stockage et de temps d'exécution des algorithmes liées à la quantité de données.

L'étude est également transposable à d'autres lieux d'échelle similaire. Les algorithmes d'analyse des données peuvent être appliqués à de nouveaux jeux de données. Quant à l'étude qualitative de la qualité des traces GPS sur le campus, des mesures physiques plus avancées pourraient être réalisées sur le principe d'une participation collective : chacun est invité à renseigner la position réelle correspondant à une de ses acquisitions GPS. Cela permettrait de quantifier les erreurs en fonction du lieu et de les prendre en compte dans les algorithmes de traitement des données.

Bibliographie

- [1] Chris Piech, *K Means*, Stanford 2013 , Disponible sur :
[<http://stanford.edu/~cpielch/cs221/handouts/kmeans.html>](http://stanford.edu/~cpielch/cs221/handouts/kmeans.html)
- [2] Étudiants de l'École Centrale de Lyon (option MD filière ADE). Rendu de mission Moovendi - Document non disponible sur internet
- [3] "Géolocalisation" [en ligne]. *Wikipédia*. Disponible sur :
[<https://fr.wikipedia.org/wiki/Géolocalisation>](https://fr.wikipedia.org/wiki/Géolocalisation) (consulté le 08/06/2018)
- [4] VAN NEERDEN Timo , "Quel est le principe de fonctionnement du GPS?" [en ligne], *Couleur-Science*, 2016, Disponible sur :
[<https://couleur-science.eu/?d=2016/06/02/17/54/23-quel-est-le-principe-de-fonctionnement-du-gps>](https://couleur-science.eu/?d=2016/06/02/17/54/23-quel-est-le-principe-de-fonctionnement-du-gps) (Consulté le 08/06/2018)
- [5] ORTEGA Luis. "Comment améliorer le signal GPS de votre smartphone Android?" *Androidpit* [en ligne]. 2018. Disponible sur :
[<https://www.androidpit.fr/comment-ameliorer-signal-gps-smartphone-android>](https://www.androidpit.fr/comment-ameliorer-signal-gps-smartphone-android)
- [6] GARNIR Henri-Pierre, STRIVAY David et BASTIN Thierry, *Le GPS et la Physique*, Science et Culture [en ligne], Novembre-Décembre 2002, Disponible sur :
[<http://randulis.free.fr/gpsdoc.pdf>](http://randulis.free.fr/gpsdoc.pdf)
- [7] *GPS: localisation et navigation* edition Hermes, Disponible sur : [<http://zebulon1er.free.fr/gps.htm>](http://zebulon1er.free.fr/gps.htm)
- [8] "K-moyennes" [en ligne]. *Wikipédia*. Disponible sur :
[<https://fr.wikipedia.org/wiki/K-moyennes>](https://fr.wikipedia.org/wiki/K-moyennes) (consulté le 05/06/2018)
- [9] GAMBS Sébastien, KILLIJIAN Marc-Olivier, NÚÑEZ DEL PRADO CORTEZ Miguel. *Next Place Prediction using Mobility Markov Chains* [en ligne] 2012, LAAS, CNRS. Disponible sur : [<http://homepages.laas.fr/mkilliji/docs/workshops/MPM12.pdf>](http://homepages.laas.fr/mkilliji/docs/workshops/MPM12.pdf)
- [10] GREGORIUS Thierry, BLEWITT Geoffrey. "The Effect of Weather Fronts on GPS Measurement", The University of Newcastle upon Tynes. *GPS World* [en ligne]. Mai 1998. p.52- 60. Disponible sur <<http://geodesy.unr.edu/publications/gpsworld.may98.pdf>>
- [11] "Location" [en ligne]. *Developers*. Disponible sur :
[<https://developer.android.com/reference/android/location/Location#getAccuracy%28%29>](https://developer.android.com/reference/android/location/Location#getAccuracy%28%29) (consulté le 08/06/2018)
- [12] "Système de positionnement en intérieur". *Wikipédia*. Disponible sur :
[<https://fr.wikipedia.org/wiki/Système_de_positionnement_en_intérieur>](https://fr.wikipedia.org/wiki/Système_de_positionnement_en_intérieur) (Consulté le 08/06/2018)

[13] Google Maps indoors, Disponible sur
<<https://www.google.com/maps/about/partners/indoormaps/>> (Consulté le 06/08/2018)

Annexes techniques

[14] Site pour tracer des trajets sur une carte : <http://mygpsfiles.com/app/>

Tableau pour la description du trajet

Paramètre	Status	Paramètre	Status
Paramètres internet		Paramètres application	
Internet activé	oui / non	App.Google Maps ouverte?	oui / non
Si internet *	wifi / données mobiles	App.s ouvertes en parallèle?	oui / non
Si données mobiles	Edge/3G/H/H+/4G	App de trackage allumée en parallèle?	oui, OruxMaps / FollowMee / non
Paramètres téléphone		Paramètres extérieures	
Type de téléphone	Androïd / iPhone	Meteo	
Version d'Androïd /iOS		Heure de la journée	
Age du téléphone + modèle		Jour de la semaine	
Niveau batterie		Nb. personnes aux alentours	

Algorithme de la méthode K-means :

```
## DESCRIPTION DU PROGRAMME :  
# Ce programme applique la méthode K-means à un ensemble de données GPS localisées  
# dans un périmètre choisi et affiche les zones obtenues.  
  
# Dans le cadre du projet, il affiche l'ensemble des données récoltées sur le campus de  
l'Ecole  
# Centrale de Lyon sur une carte du campus afin de déterminer des zones de fréquentatio  
n du  
# campus à partir des données.  
  
## PARAMETRES POUVANT ETRE CHANGÉS :  
# La liste L des points considérés  
# Le périmètre qui délimite la zone où l'on considère les points  
# datafile3 qui correspond à l'image de fond ainsi que les coordonnées (extend) associé  
es  
# à la carte  
# nombre_de_zones (jusqu'à 8 pour l'affichage avec des couleurs)  
# les couleurs des zones dans couleurs_zones  
# la graine de la fonction random.seed  
# le nombre d'itérations maximales de la méthode nbre_it  
  
## Quelques imports  
import json  
import pandas as pd  
import matplotlib.pyplot as plt  
import random as rnd  
import numpy as np  
from operator import truediv  
import time  
  
## On calcule le temps d'exécution  
t = time.time()  
  
##Créer une liste des toutes les points [latitude, longitude] de la collecte situés aux  
#alentours du campus  
L=[ "C:/Users/Maude/Documents/ECL/PE/DonneesGPS/K-means/histo_BAI_Yue.json", "C:/Users/Ma  
ude/Documents/ECL/PE/DonneesGPS/K-means/histo_Belhouari_Ahmed.json", "C:/Users/Maude/Doc  
uments/ECL/PE/DonneesGPS/K-means/histo_Chalicerne_Raphael.json", "C:/Users/Maude/Documen  
ts/ECL/PE/DonneesGPS/K-means/histo_Colas_Jules.json", "C:/Users/Maude/Documents/ECL/PE/D  
onneesGPS/K-means/histo_Fregosi_Guillaume.json", "C:/Users/Maude/Documents/ECL/PE/Donnee  
sGPS/K-means/histo_Hanser_Corentin.json", "C:/Users/Maude/Documents/ECL/PE/DonneesGPS/K-means  
/histo_Kaprelian_Theo.json", "C:/Users/Maude/Documents/ECL/PE/DonneesGPS/K-means/histo_Lyon  
_Fabien.json", "C:/Users/Maude/Documents/ECL/PE/DonneesGPS/K-means/histo_Perrine  
_Pauline.json", "C:/Users/Maude/Documents/ECL/PE/DonneesGPS/K-means/histo_Qifei_Li.json"  
]
```

```

## Définir un périmètre rectangulaire autour du campus de l'ECL (à l'aide d'Open Street Maps)
lat_nord = 45.78603
lat_sud = 45.78041
long_ouest = 4.76471
long_est = 4.77209

perimetres = [long_ouest, long_est, lat_sud, lat_nord]
#correspond aux bornes de la carte-petite-centrale : pris comme perimetre du campus

```

```

## Définir la carte de fond et son extend, le nombre de zones et leurs couleurs,
#la graine du random
datafile3 = 'C:/Users/Maude/Documents/ECL/PE/DonneesGPS/K-means/carte_centrale_petite.png'
img3 = plt.imread(datafile3)

extend3 = [4.76471,4.77209,45.78041,45.78603]

nombre_de_zones = 5 #choix du nombre de zones à calculer

couleurs_zones = ['r', 'b', 'y', 'g', 'm', 'c', 'k', 'w']
#couleurs disponibles pour les zones (8 maximum)

# utilisation de la fonction random.seed pour pouvoir retrouver certains des résultats
rnd.seed(8)

# choix d'un nombre d'itérations maximal pour limiter le temps d'exécution
nbre_it = 10

```

```

## Sélectionner les points appartenant au campus
Points = []
long, lat = [], []
# Il faut aussi sélectionner une période de temps correspondant à la collecte
for e in L:
    f=open(e)
    df = pd.read_json(f)
    donnee=df.get("locations")
    for i in range(len(donnee)):
        lo=donnee[i]["longitudeE7"]/1e7
        la=donnee[i]["latitudeE7"]/1e7
        if (lo >= perimetres[0] and lo <= perimetres[1] and la >= perimetres[2] and la <= perimetres[3]) : #points dans le périmètre
            Points.append([lo,la])
            long.append(lo)
            lat.append(la)
    f.close()

```

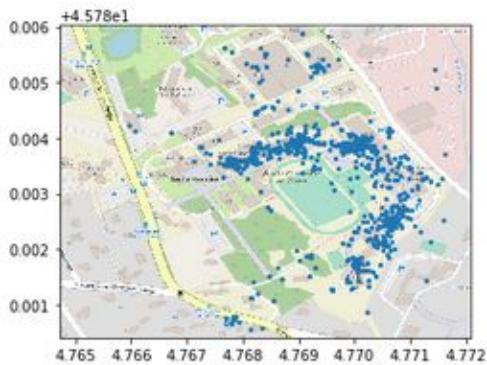
```

## Affichage des points sur une carte du campus
fig1 = plt.figure(1)

fig1
plt.scatter(long,lat, s=5, zorder=1)
plt.imshow(img3,zorder=0,extent = extend3)

plt.show()

```



```

## Délimiter un périmètre où choisir les conditions initiales des centres des zones
perimetre_centres= [min(long), max(long), min(lat), max(lat)]

```

```

## Choisir les conditions initiales de la méthode : le nombre de zones et le centre de
ces zones

# Autres types de conditions initiales : nombre de zones points distincts choisis au has
ard dans les donnees (censé améliorer la convergence)

centres_long = []
centres_lat = []
Points_donnees=Points.copy()
for i in range(nombre_de_zones) :
    r = rnd.randint(0,len(Points_donnees))
    longitude = long[r]
    latitude = lat[r]
    centres_long.append(longitude)
    centres_lat.append(latitude)
    del (Points_donnees[r])

```

```

## Affichage des centres sur la carte du campus avec un code couleur

c_long, c_lat = [], []
for i in range(nombre_de_zones) :
    c_long.append([centres_long[i]])
    c_lat.append([centres_lat[i]])

plt.figure(2)

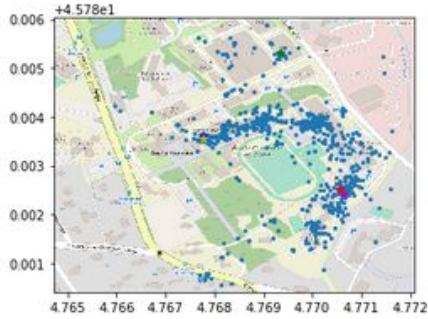
plt.imshow(img3,zorder=0,extent = extend3)

plt.scatter(long,lat, s=5, zorder=1)

for i in range(nombre_de_zones) :
    plt.scatter(centres_long[i], centres_lat[i], s=40, c=couleurs_zones[i], marker ='*'
    , zorder = 2)

plt.axis(extend3)
plt.show()

```



```

## Calcul des zones initiales

def zones_initiales(Points, nombre_de_zones, centres_long, centres_lat):
    Zones = [[] for i in range(nombre_de_zones)] #contiendra la liste des points-vecteurs appartenant à une zone d'indice i

    for i in range(len(Points)):
        long_p = Points[i][0]
        lat_p = Points[i][1]
        distances_aux_centres= []
        for j in range(nombre_de_zones):
            distance = (Points[i][0]-centres_long[j])**2+(Points[i][1]-centres_lat[j])**2
            distances_aux_centres.append(distance)
        Numero_zone = distances_aux_centres.index(min(distances_aux_centres))
        Zones[Numero_zone].append(Points[i])
        plt.scatter(long_p, lat_p, s=5, c=couleurs_zones[Numero_zone] , zorder=1)
    return Zones

## Affichage points par appartenance aux zones initiales

plt.figure(3)

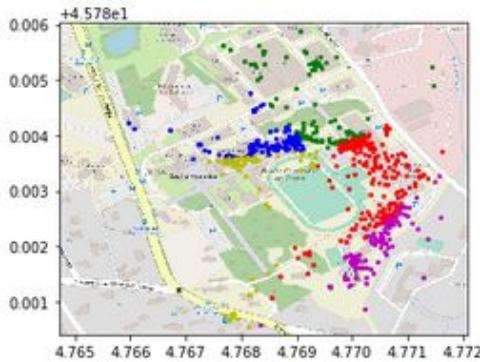
zones_init = zones_initiales(Points, nombre_de_zones, centres_long, centres_lat)

plt.imshow(img3,zorder=0,extent = extend3)

for i in range(nombre_de_zones) :
    plt.scatter(centres_long[i], centres_lat[i], s=40, c=couleurs_zones[i], marker ='*'
    , zorder = 2)

plt.axis(extend3)
plt.show()

```



```

## Calcul des 'zones denses': disques tels que si un point appartient au disque alors il est
#forcément associé au centre du disque

# Méthode adaptée à notre configuration : une forte densité de points en certaines zones

def Rayons(centres_long, centres_lat):
    # renvoie la liste des rayons des disques pour lesquel les points appartenant à un disque
    #sont forcément plus proches du centre associé que des autres centres
    rayons=[]
    for i in range(nombre_de_zones):
        distances=[]
        milieu_x, milieu_y = (centres_long[0]+centres_long[i])/2,(centres_lat[0]+centres_lat[i])/2
        distance_milieu =(centres_long[i]-milieu_x)**2+(centres_lat[i]-milieu_y)**2
        min_distances = distance_milieu
        for j in range(nombre_de_zones):
            if j!=i :
                milieu_x, milieu_y = (centres_long[j]+centres_long[i])/2,(centres_lat[j]+centres_lat[i])/2
                distance_milieu = (centres_long[i]-milieu_x)**2+(centres_lat[i]-milieu_y)**2
                distances.append(distance_milieu)
                if distance_milieu < min_distances :
                    min_distances = distance_milieu
        rayon = min_distances
        rayons.append(rayon)
    return rayons

```

```

## Calcul des zones pour une itération

def zones(former_zones, nombre_de_zones, centres_long, centres_lat, rayons): # pour n=1
    former_zones = zones initiales
    Zones = [[] for i in range(nombre_de_zones)]

    for i in range(nombre_de_zones) :
        for j in range(len(former_zones[i])):
            distance_centre = (centres_long[i]-former_zones[i][j][0])**2+(centres_lat[i]-former_zones[i][j][1])**2
            if distance_centre <= rayons[i] :
                Zones[i].append(former_zones[i][j])
                plt.scatter(former_zones[i][j][0],former_zones[i][j][1], s=5, c=couleur_s_zones[i] , zorder=1)
            else :
                distances_aux_centres = []
                for p in range(nombre_de_zones):
                    distance = (former_zones[i][j][0]-centres_long[p])**2+(former_zones[i][j][1]-centres_lat[p])**2
                    distances_aux_centres.append(distance)
                Numero_zone = distances_aux_centres.index(min(distances_aux_centres))
                Zones[Numero_zone].append(former_zones[i][j])
                plt.scatter(former_zones[i][j][0],former_zones[i][j][1], s=5, c=couleur_s_zones[Numero_zone] , zorder=1)
    return Zones

```

```

## Calcul des nouveaux centres des zones pour une étape d'itération :


```

```

def new_centres(Zones, nombre_de_zones, centres_long, centres_lat) :
    c_long, c_lat = centres_long, centres_lat
    for j in range(nombre_de_zones):
        n = len(Zones[j])
        if n!=0 :
            sum_long = 0
            sum_lat = 0
            for p in range(n):
                sum_long = sum_long + Zones[j][p][0]
                sum_lat = sum_lat + Zones[j][p][1]
            c_long[j] = truediv(sum_long,n)
            c_lat[j] = truediv(sum_lat,n)
    return c_long, c_lat

```

```

## Représentation des nouveaux centres calculés

plt.figure(4)

rayons = Rayons(centres_long, centres_lat)

Zones = zones_initiales(Points, nombre_de_zones, centres_long, centres_lat)

nvx_centres_long, nvx_centres_lat = new_centres(Zones, nombre_de_zones, centres_long, c
entres_lat)

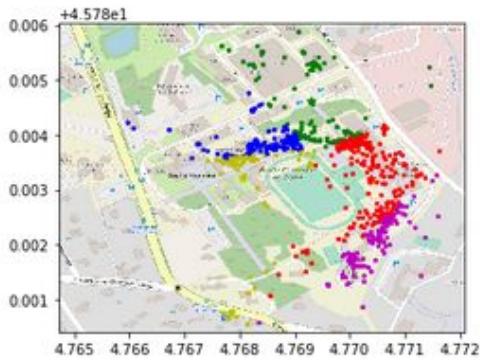
nvx_c_long, nvx_c_lat = [], []
for i in range(nombre_de_zones):
    nvx_c_long.append([nvx_centres_long[i]])
    nvx_c_lat.append([nvx_centres_lat[i]])

plt.imshow(img3,zorder=0,extent = extend3)

for i in range(nombre_de_zones) :
    plt.scatter(nvx_c_long[i], nvx_c_lat[i], s=40, c=couleurs_zones[i], marker='*', zo
rder = 2)

plt.axis(extend3)
plt.show()

```



```

## Définir un critère de fin d'itération

# Quand les centres des zones sont fixes -> boucle while ou à défaut quand le nombre
# d'itérations maximale est atteint

# On peut envisager un autre critère de convergence moins exigeant comme un faible varia
tion de
# de la position des centres.

```

```

## Programme d'itération du calcul des zones et des centres associés

def K_means(zones_initiales, nombre_de_zones, centres_long, centres_lat):
    barycentres_long, barycentres_lat = new_centres(zones_initiales, nombre_de_zones, c
entres_long, centres_lat)
    former_centres_long, former_centres_lat = [], []
    former_zones = zones_initiales
    count = 0
    while (barycentres_long != former_centres_long) or (barycentres_lat != former_centr
es_lat) or (count != nbre_it) :
        rayons = Rayons(barycentres_long, barycentres_lat)
        Zones = zones(former_zones, nombre_de_zones, barycentres_long, barycentres_lat,
rayons)
        former_zones = Zones
        c_long, c_lat = new_centres(Zones, nombre_de_zones, barycentres_long, barycentr
es_lat)
        barycentres_long, former_centres_long = c_long, barycentres_long
        barycentres_lat, former_centres_lat = c_lat, barycentres_lat
        count += 1
        print(count) #on connaît ainsi le nombre d'itérations en temps réel
    return barycentres_long, barycentres_lat, count

```

```

## Tracé de la figure de fin sur une carte du campus : centres et zones

zones_init = zones_initiales(Points, nombre_de_zones, centres_long, centres_lat)

plt.close(fig1) #destiné à effacer la figure parasite qui s'affiche avant la figure 1
# Ne fonctionne apparemment pas.

barycentres_long, barycentres_lat, compteur = K_means(zones_init, nombre_de_zones, centres_long, centres_lat)

plt.figure(5)

Zones = zones_initiales(Points, nombre_de_zones, barycentres_long, barycentres_lat)

baryc_long, baryc_lat = [], []
for i in range(nombre_de_zones):
    baryc_long.append([barycentres_long[i]])
    baryc_lat.append([barycentres_lat[i]])

plt.imshow(img3,zorder=0,extent = extend3)

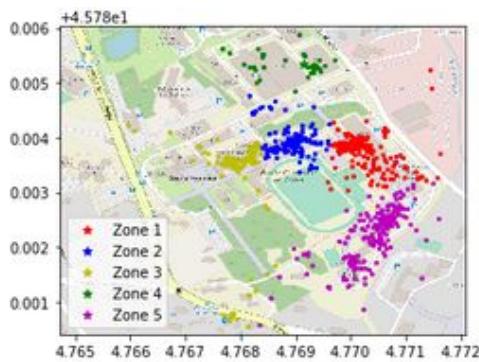
for i in range(nombre_de_zones) :
    plt.scatter(baryc_long[i], baryc_lat[i], s=40, c=couleurs_zones[i], marker='*', zorder = 2, label='Zone '+str(i+1))

plt.axis(extend3)

plt.title('Méthode K-means :\n'+ str(nombre_de_zones)+ ' zones d\'intérêt \n')
plt.legend()
plt.show()

```

1
2
3
4
5
6
7
8
9
10



```

## On vérifie le nombre d'itérations avant la convergence et on calcule le temps d'exécution
temps= time.time()-t #temps d'exécution de la méthode
print(compteur)
print(temps)

```

10
517.330518245697

Check-list de rapport de Projet d'Études

Renseigner la case par le nom du responsable, ou la date ou une simple croix lorsque la vérification a été faite.

	Vérification présence	Vérification qualité
Contenu		
Résumé en français	OK	OK
Résumé en anglais	OK	OK
Table des matières	OK	OK
Table des figures	OK	OK
Introduction	OK	OK
Conclusion générale	OK	OK
Bibliographie	OK	OK
Citation des références dans le texte	OK	OK
Forme		
Vérification orthographe	OK	OK
Pagination	OK	OK

Homogénéité de la mise en page	OK	OK
Lisibilité des figures	OK	OK