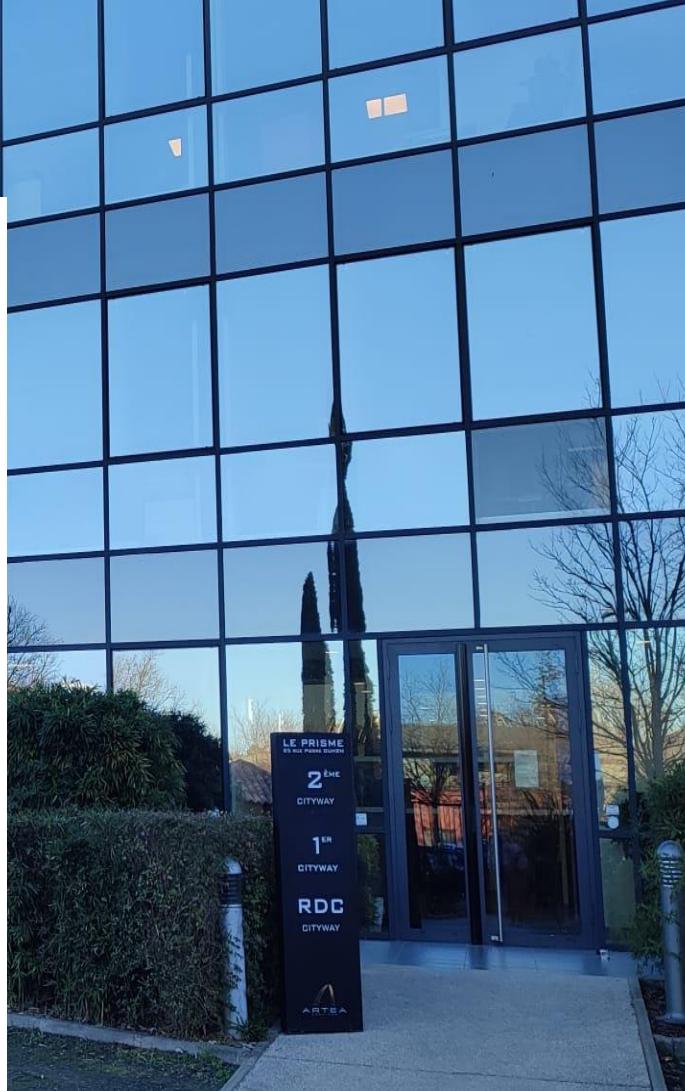


# Estimation du taux d'occupation du stationnement sur voirie à partir de données de paiement



Elise Maistre – GI03  
Université de Technologie de Compiègne  
Département Informatique

Stage assistant ingénieur TN09  
5 septembre 2022 – 17 février 2023

Suiveur UTC : M. Dritan Nace  
Tuteur de stage : M. Alexandre Iglesias  
Entreprise : Cityway  
Adresse : 85 rue Pierre Duhem 13290 Aix-en-Provence



---

# Remerciements

Je tiens à remercier toutes les personnes qui ont pu m'accorder du temps lors de mon passage à Cityway et qui ont permis la réalisation de ce stage.

Je souhaite remercier mon tuteur de stage M. Alexandre Iglesias qui m'a accompagnée durant ces six mois et m'a aidée dans l'accomplissement de ce stage, en m'aiguillant sur le chemin à suivre.

Je voudrais aussi remercier mes collègues de l'équipe TripPlanner Mme Coline Seppecher et M. Sean Shorten qui ont été très présents tout au long du stage pour me guider et m'épauler. Leur écoute et leurs suggestions m'ont été d'un très grand conseil.

J'aimerais finalement remercier M. Olivier Magnan pour ses explications diverses toujours intéressantes.

---

## Table des matières

<b>Remerciements .....</b>	<b>1</b>
<b>Résumé technique.....</b>	<b>2</b>
<b>Introduction.....</b>	<b>3</b>
<b>Présentation de Cityway.....</b>	<b>4</b>
<b>Les missions du stage .....</b>	<b>12</b>
<b>Etude du stationnement sur voirie.....</b>	<b>19</b>
Première partie : Recherches, prises d'informations et analyses.....	19
Deuxième partie : Analyse technique des données et algorithmes.....	25
Troisième partie : Résultats finaux et limites de cette étude .....	38
<b>Conclusion .....</b>	<b>51</b>
<b>Annexes .....</b>	<b>52</b>

## Résumé technique

J'ai effectué mon stage d'assistant ingénieur sur une durée de 6 mois au sein de Cityway, entreprise située à Aix-en-Provence. Cityway développe des sites internet d'information voyageur pour appréhender les réseaux de transport en milieu urbain et faciliter les déplacements. J'ai pu intégrer l'équipe du calculateur d'itinéraire (TripPlanner) en tant que développeuse.

L'objectif de ce stage est d'estimer la complexité de stationner sa voiture personnelle sur la voirie en effectuant une analyse du taux d'occupation des places de parking. A partir du taux de remplissage prédictif, il est possible de donner à l'usager un temps approximatif de recherche de place de parking.

Pour cela, j'avais à ma disposition des données de paiements provenant de la ville de Paris sur l'année 2018. Ces données proviennent des horodateurs et des applications de paiement en ligne permettant l'achat d'un titre de stationnement sur une durée précise et dans une zone spécifique.

Le stage a débuté par une phase de compréhension et de nettoyage des données. En parallèle, j'ai étudié des articles de recherche sur le sujet du stationnement en voirie. Le cœur du stage a été le travail d'analyse de données afin d'effectuer un modèle prédictif de l'occupation des places de stationnement dans les différentes zones de Paris. J'ai développé deux algorithmes de regroupement, la classification ascendante hiérarchique et le k-means, afin d'établir un modèle de jours types sur l'année. De plus, la combinaison de ces deux algorithmes a permis d'améliorer la qualité des jours types générés. Un format de données en json a été défini pour que les résultats soient exploitables par le calculateur d'itinéraires et pour que le temps prédictif de stationnement soit intégré dans les trajets en voiture personnelle. Enfin, une visualisation des résultats sur une carte de Paris a été réalisée, présentant le remplissage des places de parking sur les différentes zones en fonction du jour et de l'heure.

# Introduction

La recherche d'une place de stationnement est un évènement de la vie courante qui peut faire perdre du temps à l'usager et qui participe à l'augmentation de la pollution lorsqu'elle est difficile. Les villes tentent de limiter l'impact négatif de cette recherche en mettant en place des moyens d'informer ou de calculer le temps de stationnement selon différents critères.

Le temps passé pour rechercher une place augmente avec l'encombrement des rues et l'augmentation du nombre de voiture sur la voirie. En effet, l'augmentation des voitures présentes sur la voirie à la recherche de places accentue les embouteillages et contribue à la pollution. De plus, la part de la circulation urbaine engendrée par les véhicules en recherche de stationnement se situerait entre 5 et 10%. [17]

Il est donc utile de limiter la complexité de trouver une place de stationnement, sur la voirie, d'autant plus dans une grande ville où le temps de recherche peut-être plus long, en proposant des itinéraires alternatifs. La prise en compte du stationnement en voirie dans la recherche d'itinéraires, en pénalisant un itinéraire en voiture personnelle avec un temps de stationnement supplémentaire, pourrait influencer le conducteur à se rabattre sur une zone moins prisée pour améliorer son trajet, voire l'encourager à utiliser un moyen de transport alternatif, comme les transports en communs, pour rejoindre sa destination de façon plus attractive. Cette solution conduirait à limiter l'encombrement des rues, fluidifier le trafic tout en réduisant la pollution et pourrait contribuer à un gain de temps probable pour l'usager.

Cependant, il est difficile de mesurer correctement le nombre de véhicules à la recherche d'un stationnement. Cela nécessite des technologies capables de détecter les véhicules, par exemple, à l'aide de capteurs. [22] De plus, le calcul du taux d'occupation des places de stationnement sur voirie est une tâche beaucoup plus difficile que le contrôle du stationnement des parkings en ouvrage qui ne nécessite qu'un capteur à l'entrée et à la sortie. Dans ce rapport, nous tenterons d'estimer l'occupation des stationnements sur voirie dans la ville de Paris, non pas en utilisant des capteurs, souvent couteux et difficiles à entretenir pour les municipalités, mais plutôt en se servant des données de transactions des paiements des parcmètres et des applications.

Dans un lieu où les transports en commun ont une place centrale, comme c'est le cas pour la ville de Paris, il est intéressant de tenter l'expérience et de vérifier sa possible application. L'étude portera sur la ville de Paris, fortement exposée à ces problèmes de stationnements. Une étude poussée des données permettra d'établir une prédition pour inciter les conducteurs à choisir de nouveaux espaces pour se garer et permettre une meilleure fluidité du trafic. L'augmentation du nombre de visiteurs, ainsi que du nombre de propriétaire de véhicules aggrave le phénomène et provoque une pénurie de places de stationnement, en particulier dans certaines zones très attractives de la ville. Il est donc important de décentraliser le stationnement en proposant des itinéraires alternatifs et en dissuadant l'usager de se rendre en voiture dans les zones encombrées.

Dans ce rapport, nous tenterons de répondre à la question suivante : Est-il possible de prédire le temps de recherche d'une place de parking sur la voirie en utilisant des données de paiements ? Plusieurs défis sont à relever que ce soit concernant la fiabilité des données ou bien leur compréhension et leur analyse par des algorithmes particuliers. Il est intéressant de se pencher sur la question et de tenter d'établir un modèle de prédition sur la ville de Paris.

Nous verrons tout d'abord le fonctionnement de l'entreprise d'accueil pour la réalisation de ce projet ainsi que le service de la recherche d'itinéraire où ce projet a pu être concrétisé. Puis nous étudierons les missions de ce stage et les possibles améliorations. Par la suite, nous commencerons les recherches liées à cette étude et analyserons par une première approche les données. Ensuite, nous tenterons de créer un modèle de jours types tirer des conclusions en regroupant les jours ayant les mêmes caractéristiques à l'aide d'algorithmes de groupement. Les algorithmes détaillés seront l'algorithme de classification ascendante hiérarchique (CAH) [22] et l'algorithme k-means. [18] Finalement, nous ajouterons des données complémentaires pour étoffer les résultats et tirer une prédition plus juste du temps de recherche de stationnement sur la voirie. Ainsi, nous résumerons ce que travail d'analyse de données a apporté à ce projet de prise en compte du stationnement sur voirie. Il sera également intéressant de s'interroger sur les limites de cette analyse et sur les éventuelles extensions de ce travail.

# Présentation de Cityway

Cityway est l'entreprise dans laquelle j'ai été accueillie pour y réaliser mon stage. Cette immersion au sein de son organisation et plus précisément au sein de l'équipe du calculateur d'itinéraires (équipe également appelée *Trip Planner*), constitue une approche épanouissante pour moi et une découverte de la vie en entreprise.

## 1) Un peu d'histoire

L'entreprise Cityway a été créée en 2001 par Mr Laurent Briant pour répondre à un besoin d'information des voyageurs et simplifier la mobilité de chacun au sein des réseaux de transport urbains. Cityway est une filiale du groupe Transdev et, avant 2011, de Véolia Transport. L'entreprise se positionne sur les marchés publics d'information voyageur auprès des administrations des villes et métropoles. Un responsable de transition, M.



François Barraud, issu de Transdev depuis mi-2022, est en train de réorganiser la stratégie commerciale et technique de l'entreprise, afin de répondre à de nouveaux projets dans de meilleures conditions financières. Cette transition a pour but, en changeant l'organisation interne de la gestion des projets auprès des différentes villes, de gagner plus de stabilité sur le marché du transport et de l'information voyageur et de s'étendre à l'international

Le siège de Cityway est situé à Aix-en-Provence. Au total, Cityway compte 130 employés répartis dans 11 agences en France et dans le monde. La plus grande partie des employés sont au siège d'Aix-en-Provence.

Les solutions développées par l'entreprise sont présentes sur tout le territoire français (Beauvais, Dijon, Lyon, Paris...) et également présente au Canada et aux Etats-Unis.

Depuis quelques années, l'entreprise reçoit des prix de l'innovation, par exemple en 2018 elle a été lauréate du prix French Mobility pour son Système Intégré de Services à la mobilité de l'Oise. (SISMO)

## 2) Sa mission

Son objectif est de proposer à différentes collectivités comme la Métropole d'Aix-Marseille, l'Île de France Mobilité, la Bourgogne Franche Comté ou encore la Nouvelle Calédonie ou Toronto, des solutions numériques pour simplifier, optimiser et piloter la mobilité sur leur territoire et notamment de mettre en avant leurs offres de transports en communs et leur offre de mobilité dite « douce » (modes de transport non polluants comme le vélo en libre-service).

La mission de Cityway est donc de proposer un service numérique aux collectivités pour informer les voyageurs, proposer des trajets multimodaux prenant en compte plusieurs types d'offres ainsi que de faciliter la vente des titres de transports. Elle propose également un service de contact, pour les usagers sous forme de centrales d'appels.

Cityway veut permettre l'accès à la mobilité pour tous, d'abord grâce à un système d'information multimodal, c'est-à-dire intégrant tous les modes de transports et types de données liées à ces modes, ensuite en permettant le paramétrage d'options spécifiques comme l'accessibilité pour les voyageurs à mobilité réduite, ou encore la volonté de garder son vélo dans les transports en commun (véhicules spécifiques pour le « vélo à bord »). Ainsi, les solutions proposées par Cityway sont d'une grande diversité, avec par exemple l'utilisation de la trottinette, l'indication du nombre de vélos en libre-service dans les stations en temps réel ou encore le transport à la demande dans les zones peu denses. Cityway travaille également en partenariat avec des prestataires externes, que ce soit pour du covoiturage (BlaBlaCar, BlaBlaCar Daily, Karos, Klaxit...) ou pour des correspondances transfrontalières (projet EU-Spirit pour interfaçer la région de Grand-Est avec l'Allemagne, le Luxembourg et la Suisse). Néanmoins, le cœur de métier de l'entreprise reste l'information liée aux transports en commun (bus, métro, tram, bateau...) avec les horaires en temps réel des différents transports et les perturbations (travaux, grèves,

incidents...). L'équipe calculateur d'itinéraires intègre également les cartographies piétonnes, vélo et voiture avec les spécificités liées à ces modes : ajustement des vitesses de marche à pied et de vélo en fonction du dénivelé, trafic routier en temps réel, perturbations routières...

### 3) Ses produits

Cityway crée des produits et services pour pouvoir gérer le MaaS (Mobility as a Service) qui constitue sa solution pour toutes les politiques de mobilités. La marque combinant les différents produits se nomme Manett et est constitué des 5 points clés de Cityway. C'est-à-dire la possibilité de réservation, l'information voyageurs, l'optimisation d'un réseau de transport, l'intégration de données externes et la vente de titres de transport.

Actuellement, ses produits sont déployés dans plus de 150 villes à travers le monde, principalement en France. Les solutions apportées par Cityway se présentent sous la forme d'applications et de sites internet pour permettre aux utilisateurs de s'informer sur les disponibilités tout en optimisant leur déplacement et en leur proposant différentes alternatives. Il est également possible d'acheter des titres de transport directement sur les sites web ou les applications mobiles. Les points clés de ce service correspondent à la multimodalité, c'est-à-dire à la proposition d'itinéraire combinant différents modes de déplacements. Le but de Cityway étant de rester l'acteur le plus innovant de ce secteur.

L'objectif de Cityway est de fluidifier tous les services liés à la mobilité au sein d'une seule application, cela passe donc par des exigences sur ses produits rendus. Ainsi ses produits ont pour vocation d'agrégner tous types de services, que ce soit les différentes offres de transport sur le territoire avec des visualisations cartographiques mais aussi de permettre l'utilisation d'un calculateur d'itinéraire, en passant par la prise en compte d'informations temps réel, des perturbations, des déviations, des changements d'horaire, tout en informant les voyageurs par des notifications sur le trajet et en permettant également l'achat des titres de transport directement depuis l'application.

Cityway travaille avec les collectivités locales, les réseaux de transports et les grands acteurs publics (Ministère des transports, ADEME, ...) pour réaliser ses produits.

Sa clientèle est majoritairement constituée des collectivités locales ainsi que des exploitants du secteur transport. Les collectivités, qui sont des métropoles et des agglomérations représentent 2/3 du chiffre d'affaires de Cityway. Les clients font alors des appels d'offre et plusieurs entreprises répondent à ceux-ci, pour tenter de remporter le marché. En cas de succès, les équipes de Cityway commencent la réalisation du système basé sur son produit tout en écoutant le client pour répondre à ses attentes spécifiques.

Les partenaires de Cityway sont divers mais permettent l'apport de données externes pour certains types de transports. Par exemple, des données venant d'acteurs de covoiturage comme Blablacar ou Klaxit ou bien des données spécifiques à l'utilisation du vélo et des pistes cyclables comme Géovélo, leader dans son domaine. L'entreprise travaille également sur certains projets avec les pays frontaliers pour rattacher l'offre de transport, avec celle d'un autre pays, ce qui représente un défi d'une autre envergure. Par exemple, avec l'Allemagne dans son projet pour Grand-Est.

Les concurrents de Cityway sont de deux types. Tout d'abord, les concurrents directs qui répondent aux appels d'offres des clients et qui développent des solutions informatiques de mobilité similaires à celles de Cityway, comme Instant System.

L'autre catégorie de concurrents sont les concurrents indirects qui développent leurs propres solutions directement pour le grand public comme Move IT, City Mapper ou encore Google Maps. Mais dans le cas de ces concurrents indirects, n'ayant pas de liens directs avec les clients, leurs données peuvent parfois être incomplètes.

#### 4) L'évolution de sa stratégie

L'ancienne stratégie de Cityway reposait sur le produit vendu et des difficultés de maintenance se posaient. En effet, une fois le marché remporté, le produit était entièrement conçu sur-mesure selon les besoins et les volontés du client, durant 2 ou 3 ans. Puis, une fois le produit fini, peu de suivi était réalisé. Ainsi comme le produit reste environ 5 ans sur le marché, au bout de quelques années, en cas de problème, la maintenance devenait très difficile.

La nouvelle stratégie de Cityway est d'établir un CityOne ou MaaS France, c'est-à-dire centraliser tous les projets français sur un même serveur, de regrouper tous les cas et de proposer une seule offre sur une partie du territoire. Le projet consiste à créer un socle unique sur toute la France pour y déployer des solutions numériques génériques pour différents clients. L'avantage serait de minimiser le coût de l'intégration de nouveaux clients sur le territoire français et d'être plus rapide dans la réalisation du projet. De plus, la maintenance serait plus facile à suivre et à réaliser car tout serait regroupé. CityOne est un projet ambitieux de par la taille du territoire à couvrir, le volume de données à intégrer, et les contraintes de performances à respecter (temps de réponse des différentes applications vendues). Le projet est également de centraliser tous les serveurs et de les faire héberger par des services externes qui prendront en charge leur maintenance, ce qui commence déjà à être le cas.

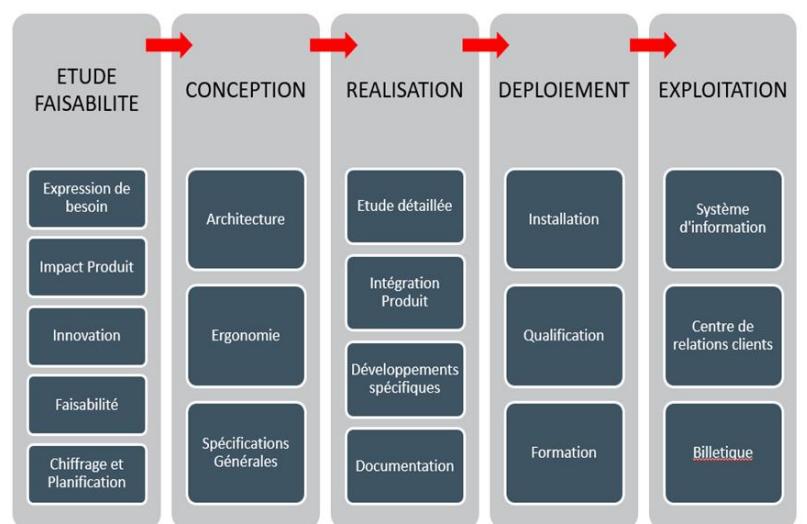
Cette nouvelle stratégie passe par la restructuration interne des équipes et des postes notamment avec la création d'une équipe uniquement prévue pour le MaaS France.

L'autre stratégie de Cityway est de s'imposer en France et de gagner de plus en plus de marchés à l'étranger et notamment là où elle est déjà implantée, au Canada et aux Etats-Unis. La stratégie est donc d'être plus présent sur les territoires d'Amérique du Nord et d'Europe.

Ainsi Le service que Cityway propose a pour but d'être compétitif, facile d'utilisation et durable.

#### 5) Son processus

Le processus actuel après avoir remporté un projet sur le marché est de, tout d'abord, réfléchir aux moyens de mise en place pour réaliser l'étude de faisabilité, c'est-à-dire comprendre et trouver les éventuelles difficultés qui pourront se poser pour réaliser les besoins du client. La phase suivante est celle de la conception, c'est-à-dire la pré-formalisation du sujet, et l'architecture du projet. Ensuite, se déroule la phase de réalisation, c'est dans cette partie que les développements techniques sont réalisés. Lorsque le projet a été testé en préproduction et qu'il est opérationnel il peut être déployé en production, c'est à dire accessible à tous les usagers. A la suite de quoi se déroule une phase d'exploitation durant laquelle des opérations de maintenance devront être réalisées.



**Figure 2 : Format du processus de mise en production d'un projet**

## 6) Son organisation

L'entreprise est découpée en 3 pôles. Le premier concerne l'administratif, c'est-à-dire les ressources humaines, comptabilité, financier et administratif. Le second est le pôle commercial avec le marketing et l'avant-vente.

Le dernier pôle est le pôle technique et est le plus grand des trois. Il est lui-même découpé en cinq parties. Une partie concernant l'interface et l'expérience utilisateurs, qui comprend la création du design de l'application. Une autre partie de conception et innovation, qui sont réalisées par les responsables produits qui ont pour mission de comprendre les besoins des clients pour les retrancrire techniquement et pouvoir les chiffrer. Les chefs de projet qui gèrent le projet au cours de son développement jusqu'à la livraison tout en étant en interaction avec les clients et les équipes techniques. Les responsables d'exploitation qui réalisent la maintenance des projets, une fois que les projets ont été livrés et qui dispose aussi d'une centrale d'appel qui répond aux voyageurs. La dernière partie est la partie la plus technique, composée des développeurs logiciels et de plusieurs équipes qui interagissent beaucoup entre elles pour créer les produits des différents projets. C'est la partie dans laquelle j'ai été accueillie pendant ces 6 mois.

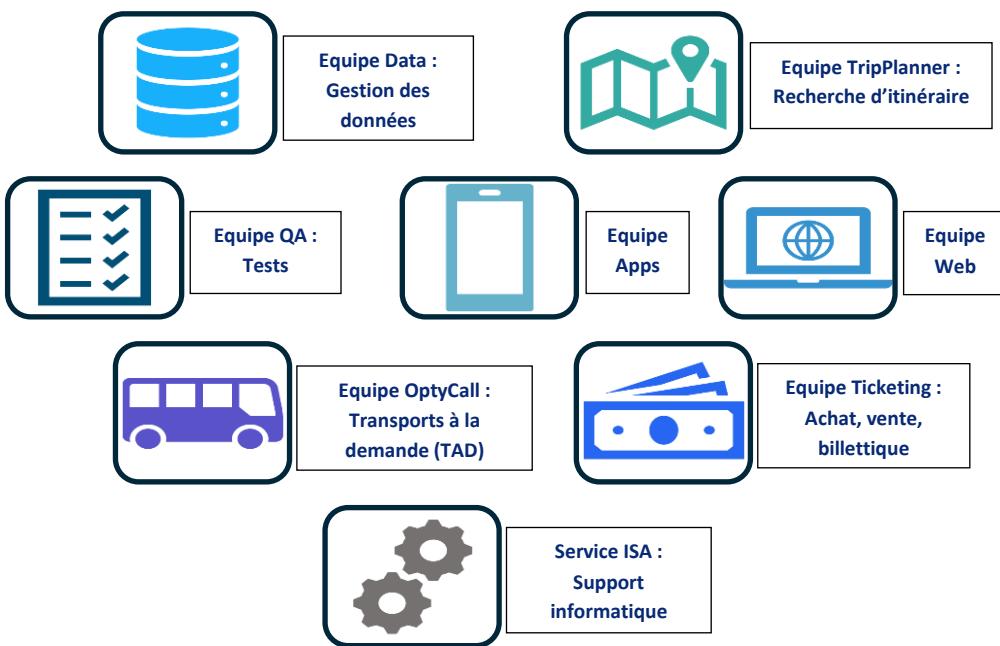


Figure 3 : Zoom sur la partie techniques et les différentes équipes qui s'y trouvent

Il existe également un service de support informatique en charge de la sécurité et de l'administration système et des réseaux. Ainsi qu'une partie de Business Intelligence, c'est-à-dire la formation de rapport statistiques concernant les exploitations des clients.

La partie technique est composée de 7 équipes :

- L'équipe « Data » s'occupe de l'intégration des données externes et de la gestion des bases de données nécessaires au bon fonctionnement des solutions comme les données cartographiques ;
- L'équipe « Trip-Planner ou Recherche d'itinéraire » est chargée de développer le calculateur d'itinéraire ;
- L'équipe « QA : Quality Assurance » teste les différentes solutions à la recherche de « bugs » pour assurer leur bon fonctionnement ;
- L'équipe « Apps » gère le développement des solutions sur téléphones (pour les systèmes d'exploitation Android et IOS) ;
- L'équipe « Web » réalise le développement des sites internet ;
- L'équipe « Optycall » travaille sur le transport à la demande ;
- L'équipe « Ticketing » est responsable de la partie vente du site
- L'équipe « ISA ou support informatique » gère le matériel et les logiciels utilisés par la société ;

## 7) L'Equipe TripPlanner

L'équipe TripPlanner est l'équipe que j'ai eu la chance d'intégrer durant mes 6 mois de stage.

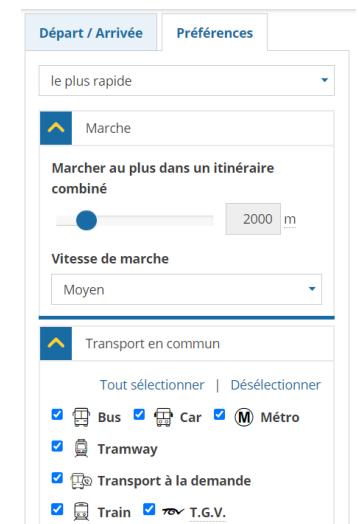
- **Sa mission**

Elle a pour mission principale de développer le calculateur d'itinéraire de Cityway. C'est-à-dire de renvoyer les meilleurs résultats de recherche d'itinéraire. A partir de requêtes de RI elle doit renvoyer des solutions multimodales. Ces solutions prennent en compte différentes données de transports en commun, les lignes et les interconnexions existantes entre ces lignes. Que ce soit des bus, des métros, des trains ou des trams avec des lignes spécifiées et des horaires de passage. Mais également la prise en compte des trajets routiers, avec la marche, le vélo ou la voiture qui dépendent de la cartographie. De plus, le prix des itinéraires doit être calculé et renvoyé par cette équipe.

Le point central du système est donc le calculateur d'itinéraire qui permet le retour d'itinéraires cartographiés.

- **Le calculateur d'itinéraire**

Le calculateur d'itinéraire de Cityway est **multimodal**, c'est-à-dire qu'il peut prendre en compte plusieurs moyens de transports différents, il peut renvoyer plusieurs itinéraires pour une même demande avec des modes de transport différents. Il est de plus particulièrement configurable, pour décider des contraintes sur certains modes de transports à privilégier selon les demandes des clients et des usagers directement sur l'application. En effet, certains réseaux demandent de favoriser des moyens alternatifs à l'usage de la voiture. De plus, sur l'application ou le site internet il est possible de réaliser plusieurs types de configuration. Il est possible de définir des critères d'optimisation comme par exemple la vitesse de marche, ou le nombre de changement, les pistes cyclables. Mais il est également possible de définir des contraintes de résultats. Par exemple l'utilisateur peut définir la distance maximale de marche à pied proposée, de même pour le vélo ou la voiture, ou encore des restrictions sur certains modes de transports et même l'interdiction de certains réseaux de transports pour faire correspondre les propositions à son abonnement. Toutes ces caractéristiques sont enregistrées directement dans le calculateur d'itinéraire et prises en compte lors de la recherche de la solution la plus adaptée à l'usager.



*Figure 4 : Sélection des contraintes préférentielles par l'usager sur le site Fluo*

Par ailleurs, le calculateur d'itinéraire de Cityway est **intermodal** c'est-à-dire qu'il peut combiner différents modes de transports dont les modes externes avec l'aide des API externes comme les acteurs de covoiturage notamment. Le transport à la demande est également utilisé.

Comme on le voit avec l'image ci-contre extraite du calculateur en production fluo sur la région Grand-Est, les itinéraires proposés sont multimodaux, avec des propositions de transports en commun, de voiture, vélo et marche. Ils sont également intermodaux, par exemple ici ils allient la marche, le métro et différents trains pour arriver à destination.

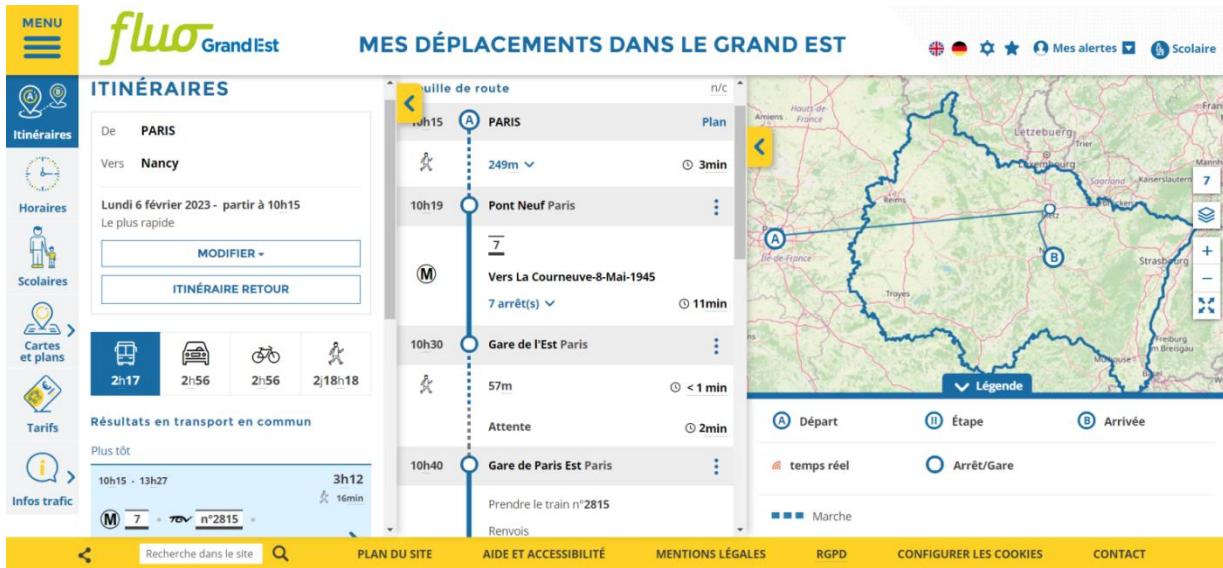


Figure 5 : Extrait du site Fluo, calcul d'itinéraire entre Paris et Nancy à 10h15 (solution transport en commun)

Le calculateur d'itinéraire utilise différents algorithmes pour son fonctionnement. Il reçoit en données d'entrée les adresses de départ et d'arrivée ainsi qu'une heure, celle de départ ou d'arrivée. A partir de ces données il calcule les différentes possibilités en explorant les combinaisons et cela en utilisant deux types d'algorithmes. Pour tout ce qui concerne les parcours de graphes routiers (marche à pied, vélo ou voiture) le calcul est fait avec l'algorithme A\*. Les intersections représentent les noeuds du graphe et les routes sont représentées par les arrêts. Ce graphe est orienté et chacune des arrêts possède un poids dépendant de plusieurs critères. Pour les transports en commun c'est l'algorithme de Djikstra qui est utilisé. L'objectif est donc de trouver le plus court chemin dans un graphe. La différence avec le routier c'est que des horaires de passages sont à prendre en compte dans le graphe. L'équipe travaille actuellement sur un nouveau projet pour améliorer le calculateur et l'espace mémoire utilisé. L'objectif est de remplacer l'algorithme de Djikstra par le Connection Scan Algorithm (CSA), qui changerait la manière dont est calculé l'itinéraire. En effet, il serait plus performant car les données explorées seraient continues en mémoire (liste des horaires) à la différence d'un graphe qui suit la topologie du réseau. De plus, différents paramètres sont à prendre en compte lors des connexions entre les différents transports. Que ce soit le temps de correspondance au sein d'un même arrêt, le temps de montée et de descente pour un bus ou encore les horaires en temps réel et les perturbations routières ou pas. De plus certains modes de transports doivent être pénalisé ou certains réseaux notamment si le prix est beaucoup plus élevé pour une faible différence. Il est également important de proposer de la diversité, en général ce sont 3 itinéraires qui sont proposés : le plus optimal et 2 autres cherchant un compromis entre la qualité de la proposition et la variation des lignes ou des réseaux de transports.

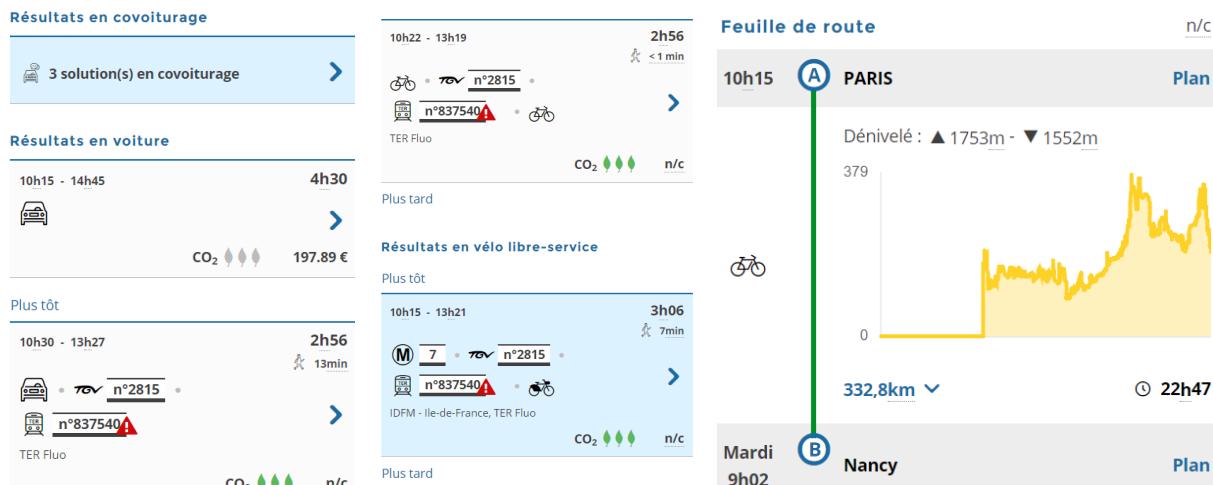


Figure 6 : Extrait du site Fluo, calcul d'itinéraire entre Paris et Nancy à 10h15 (solution voiture et vélo)

Ci-dessus sont illustrés les différents itinéraires concernant la recherche de solution en voiture et en vélo. On voit que des propositions de covoiturage sont retournées. De plus une proposition alternative à la voiture est proposée, cette solution étant intermodale. La solution suivante concerne l'utilisation du vélo. La proposition à droite indique le dénivelé (les données de dénivellées n'étant importées que pour la région Grand-Est) mais au centre nous pouvons voir des solutions alternatives comme l'utilisation du vélo au début et à l'arrivée avec transport du vélo à bord ou bien l'utilisation de vélo en libre-service à l'arrivée.

- Son fonctionnement technique

Les requêtes de recherche d'itinéraires sont reçues via la partie médias qui représente les applications ou site en ligne liée à une recherche d'un utilisateur. Elles peuvent également être reçues via des utilisateurs externes dans le cas où les projets sont divisées entre plusieurs entreprises qui répondent à l'appel d'offre, lorsque le calculateur d'itinéraire est le seul utilisé dans le projet et que d'autres entreprises s'occupent des autres parties du projet.

Cette requête passe par la politique de mobilité qui réalise les calculs selon la requête et les filtres. Cette politique de mobilité est située dans le JPS pour Journey Planner Service au centre du calcul. Il sert

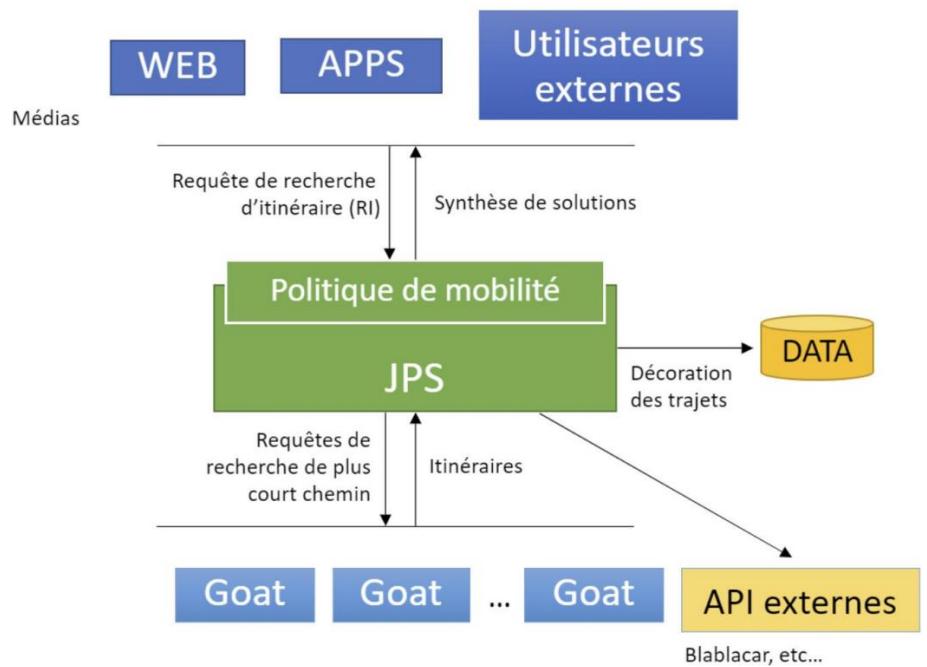


Figure 7 : Fonctionnement technique du calculateur d'itinéraire à Cityway

de liens entre les différentes informations et permet d'effectuer le retour d'une solution cohérente et compréhensible. Le JPS transmet donc la requête RI aux moteurs, que l'on appelle « Goat » (pour Go at : aller à), et qui ont pour mission de calculer l'itinéraire du plus court chemin, de la manière décrite ci-dessus. Le JPS découpe les RI combinant plusieurs modes en RI avec un seul mode, car Goat n'accepte pas de RI combinés. Les Goat renvoient donc un itinéraire au JPS, qui s'occupe ensuite de la décoration de trajet. Il est important de comprendre que les Goat doivent calculer un itinéraire en quelques centaines de millisecondes afin de proposer un trajet fluide à l'utilisateur. Pour cette raison, on utilise des cartes sous forme de graphes ne possédant que le minimum d'information nécessaire pour le calcul afin de pouvoir optimiser la vitesse de calcul. Chaque route, dans le graphe utilisé par les Goat, se voit ainsi attribuer un identifiant, permettant de faire le lien entre une carte complète et sa version grappe. Il est donc nécessaire, une fois l'itinéraire défini, de réassocier le grappe à la carte, afin d'afficher toutes les informations inutiles au calcul d'itinéraire mais utiles à l'utilisateur, comme le nom des rues, des lignes de bus, des arrêts, et ainsi de suite. Toutes ces informations de dégradations ainsi que les graphes sont stockées dans les bases de données de Cityway. Le JPS peut également, en fonction des projets, faire appel à des API externes comme la SNCF ou Blablacar en complément de la décoration pour afficher d'autres informations comme les trajets disponibles en covoiturage ou le prix de trajet de train. Une fois décorée, la solution est renvoyée au média afin de pouvoir s'afficher à l'utilisateur.

- Interactions

L'équipe TripPlanner se trouve dans le pôle technique, elle interagit en continu avec les autres intervenants de ce pôle que ce soit les chefs de projets avec qui elle aide à la création de spécifications techniques qui servent de formalisation des besoins du client. Ou même avec les clients eux-mêmes pour spécifier des points techniques concernant le calculateur d'itinéraires et l'ajout de nouveaux modes ou fonctionnalités en son sein. Elle interagit

également en cas de maintenance sur d'ancien projets en phase de production, si des problèmes sont reçus. De plus, elle interagit en permanence avec les autres équipes techniques pour développer les produits et continuer l'élaboration de son calculateur.

Ces interactions peuvent avoir lieux sous forme directe, c'est-à-dire face à face, de manière orale ou lors de réunion ou bien sous forme de visioconférence ce qui est souvent le cas lors d'interaction avec des clients ou des partenaires ou en cas de télétravail. Elles peuvent également avoir lieux sous forme indirecte ou écrite avec l'utilisation de mails, de teams ou bien de tickets Jira, qui sont envoyés et reçus par des équipes différentes et concerne des requêtes, des résultats ou des problèmes signalés et occupe une grande partie du temps de travail des équipes techniques.

- Organisation de l'équipe

L'équipe est composée de trois personnes, un chef et deux développeurs qui ont plusieurs objectifs.

Ils doivent tout d'abord traiter les tickets Jira reçu, ils doivent également être présent lors de la phase de conception des produits concernant le calculateur d'itinéraire mais également lors de la phase de maintenance des anciens projets. Leur deuxième objectif est d'améliorer et de maintenir leur calculateur d'itinéraire JPS et Goat. Les outils utilisés sont divers que ce soit Git pour partager en permanence leur code ou bien sharepoint en cas de présentation ou de spécifications, tout ce qui est document avec les clients, mais également répondre aux tickets Jira ou bien valider leurs travaux à partir d'imputations sur les différents projets travaillé chaque jour ou partie de journées. Ils peuvent également se servir de l'outil Grafana permettant d'établir des statistiques sur les projets en production.

Par ailleurs, le calculateur d'itinéraire se décompose en une API Web développée en C# et un moteur en C++. L'environnement de développement est Visual Studio 2019 sous Windows, avec un stockage de données Microsoft SQL Server.

Leur méthode est donc d'établir un planning avec leur chef et de réaliser certaines réunions au début de l'année ou de nouveau projets pour définir les tâches de chacun qui seront évolutives. La méthode est plutôt de suivre les projets les besoins au cours du temps plutôt que de figer un planning fixe qui ne sera pas représentatif des futurs besoins. La motivation est donc un facteur indispensable au sein de cette équipe.

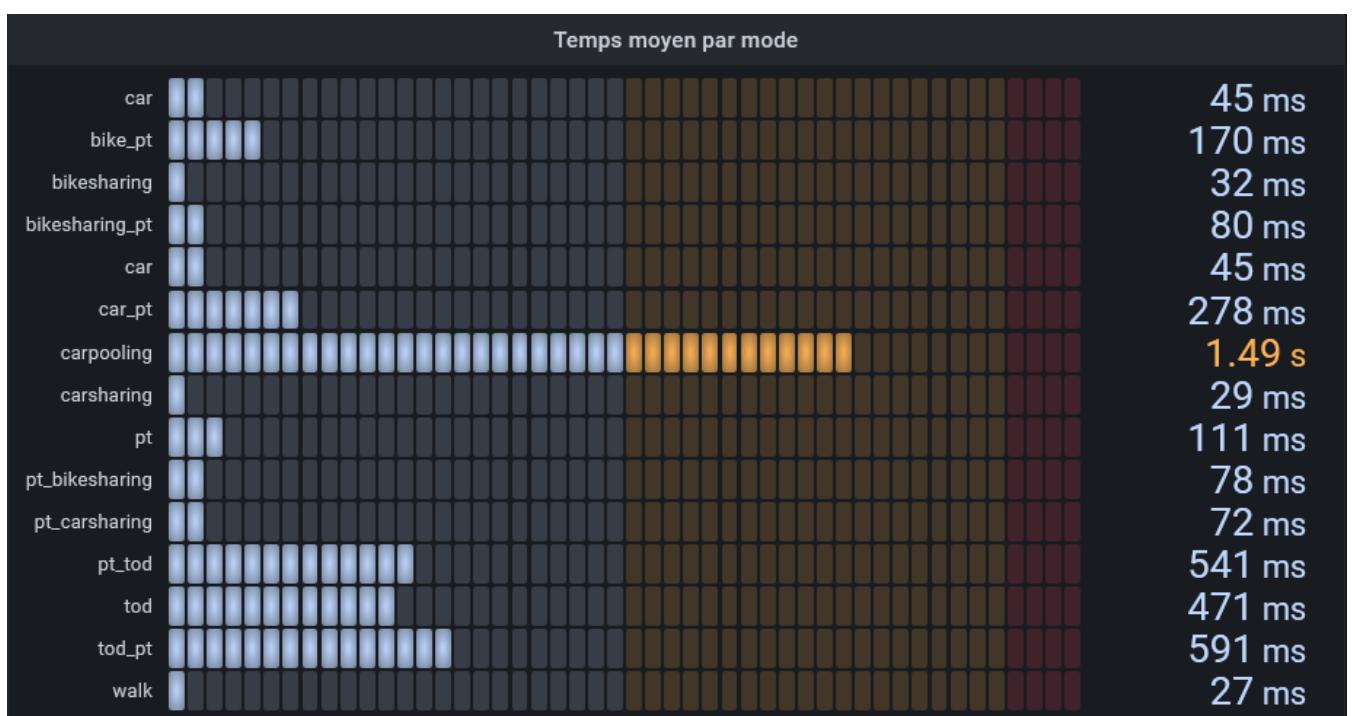


Figure 8 : Temps de réponse moyen selon les modes cherchés par le calculateur d'itinéraire, outil : Grafana

# Les missions du stage

L'objectif de ce stage est d'estimer l'occupation des places de parking sur voirie, pour une date donnée et une heure donnée, à partir de données historisées. Plusieurs fonctionnalités intéressantes pourraient en découler :

- Informer l'usager sur la facilité de se garer dans une zone géographique en fonction de la date et de l'heure de la journée
- Dans les itinéraires en voiture personnelle proposés par la recherche d'itinéraire, évaluer le temps nécessaire pour garer son véhicule à proximité de la destination
- Proposer des itinéraires alternatifs lorsque la zone géographique demandée est trop occupée, en proposant une transition vers les transports en commun dans une zone moins tendue.

Le problème d'un tel projet est la difficulté d'obtenir des données de terrain. En effet, l'utilisation de capteurs au sol pour évaluer l'occupation des places de parkings et en tirer modèle prédictif de remplissage est un dispositif complexe et donc couteux à mettre en place. Bien qu'extrêmement précis, ce dispositif est difficile à installer sur une zone géographique étendue et pour une durée suffisamment importante. En pratique, ce type de données est donc très rare et épars.

Ainsi, l'enjeu principal de ce projet est d'évaluer la faisabilité d'une estimation de l'occupation des places de parkings en voirie à partir des données de paiement des usagers.

Nous verrons dans une première partie l'énoncé initial du sujet et son évolution avec le temps en fonction des premiers résultats obtenus et des différents problèmes rencontrés. Nous nous pencherons donc sur le planning initial et sur ses changements, puis sur la réalisation de ce projet et sur les contributions apportées. Finalement, nous étudierons quelles pourraient être les extensions de ce travail.

## 1) Le sujet

Le sujet a légèrement évolué au cours du stage, après les premiers résultats obtenus et les premières recherches réalisées. Le sujet initialement proposé consistait au « développement de la prise en compte du stationnement en voirie, et des disponibilités temps réel ou prédictives de places libres ». Il s'est avéré dès le départ que la possibilité d'obtenir des données temps réel était exclue, les seules données disponibles étant les données des transactions de stationnement de 2018. L'utilisation de données en temps réel pour compléter un modèle prédictif basé sur les données historisées de 2018 pourrait compléter le résultat de ce stage et améliorer la précision des résultats dans le futur. Cependant, les données de paiement en temps réel restent très difficiles à obtenir car elles nécessitent un partenariat avec les sociétés de gestion du stationnement ainsi que la mise en place d'un flux direct de données temps réel.

Les différents objectifs définis aux départs ont été introduits et respectés dans leur ensemble. Voici ci-dessous la manière dont ils ont été réalisés.

Objectif initialement présentés	Réalisation des objectifs
<b>Compréhension du fonctionnement du calculateur d'itinéraire de Cityway et découverte des différents aspects métiers liés à la mobilité.</b>	Découverte de l'entreprise, des liens entre les pôles techniques et plus spécifiquement de l'équipe TripPlanner et son calculateur d'itinéraire, présenté dans la première partie de ce rapport
<b>Elaboration d'un plan de développement avec relecture et mise à jour éventuelle de spécifications fonctionnelles. En collaboration avec l'équipe Trip Planner.</b>	Recherches réalisées sur le sujet et lecture de la spécification fonctionnelle, puis réflexion sur la suite du projet et établissement du plan d'action et du déroulement. On peut suivre le plan de développement via le planning GANTT final dans la section suivante.
<b>Rédaction de spécifications techniques.</b>	Rédaction de rapports intermédiaires, techniques sur l'avancement du travail pour suivre le processus et

	notamment formalisation des algorithmes nécessaires à la prédiction.
<b>Développement d'un modèle relationnel de stockage des données en collaboration avec les équipes Data, Archi et Trip Planner.</b>	Stockage des données de prédiction selon les jours types formalisés sous format json. Explication plus détaillée dans la dernière partie de ce rapport.
<b>Développement d'un outil de simulation de données (géolocalisation des places, temps réel...).</b>	N'ayant pas eu de données temps réel, il n'a pas été possible de le prendre en compte. La visualisation des données est présentée sous forme d'une carte visuelle, modifiable dans le temps selon la date demandée.
<b>Tests, corrections.</b>	Débogage et modifications. La vérification des données est difficile étant donné que les réalisés de terrains n'ont pas été enregistrés en 2018.
<b>Rédaction de rapport de stage/résultats.</b>	Rédaction du rapport ci-contre.

L'objectif principal de ce stage consiste en la production des données prédictives d'occupation des places de parking sur voirie. Ma mission a été de trouver et de développer les moyens techniques d'analyse de données et les algorithmes afin de produire le modèle prédictif attendu. Le sujet de ce stage a donc été d'analyser les données de paiements reçues dans le cadre d'un projet avec la ville de Paris et d'en tirer un schéma de disponibilité de la voirie, utilisable directement par le calculateur d'itinéraire. Ce schéma de disponibilité doit indiquer, avec une certaine précision, le temps de recherche d'une place de parking libre selon la date, l'heure et le lieu. Ainsi, à partir des données de transactions issues de l'année 2018, il faut prédire sur les années futures l'attractivité de certaines zones géographiques ainsi que la fluctuation des places occupées par rapport à l'heure de la journée et au type du jour concerné (vacances scolaires ou non, par exemple).

Le travail de prise en compte du stationnement sur voirie dans le calculateur d'itinéraire avait déjà été abordé en 2017, après la demande de la ville de Paris dans le cadre du projet Mi2. Une spécification fonctionnelle avait été réalisée, avant la réception des données fournies par la ville de Paris, en février 2017. Cette pré-étude considère les différentes possibilités de format des données de stationnement : données temps réel, prédictives ou historisées. Ce projet avait été laissé à ce stade depuis 2017. Les données historisées ont ensuite été reçues, retracant l'historique de l'année 2018. Elles se décomposent en 3 jeux de données explicités plus en détail dans la partie dédiée à l'étude du projet.

Cette spécification fonctionnelle, ainsi que les jeux de données des paiements effectués sur les places de stationnement à Paris sont les points de départs de ce stage.

Les objectifs de ce stage en termes d'applications fonctionnelles des données générées étaient, d'une part la prise en compte d'un temps de stationnement prédictif dans le calculateur d'itinéraires, et d'autre part l'information voyageur des zones tendues, par exemple en proposant un résultat visuel sur une carte de Paris.

L'objectif concernant le développement de la prise en compte des résultats dans le calculateur d'itinéraire n'a pas été réalisé car jugé non pertinent sur la fin du projet. En effet, la partie technique concernant l'analyse des caractéristiques des jours types pour établir le modèle prédictif s'est avérée plus longue que prévue. De plus, le projet Mi2 a été clos et aucun autre projet en cours n'avait besoin de cette fonctionnalité.

Cet objectif a donc été restreint à la formalisation des résultats en un modèle json, afin de pouvoir réutiliser ces résultats dans le futur sur de nouveaux territoires et avec de nouvelles sources de données. Ce format générique pour décrire le taux de remplissage de zones géographiques par jours types a été institué et pourra servir de base aux futures applications dans le calculateur d'itinéraires.

D'autre part, la seconde application de ce projet a été développée avec la mise en place d'une carte temporelle permettant de visualiser les différents niveaux de complexité du parking sur voirie dans les zones de Paris. Six catégories ont été définies pour décrire la difficulté du stationnement : vide / négligeable / court / long / très long / extrêmement difficile.

Quant au déroulement du stage, il n'était pas défini précisément au départ et les applications finales dépendaient de la suite de mes recherches. Il fallait d'abord m'approprier les données avant de pouvoir commencer

à réaliser une prédition pour le futur. J'ai donc contribué, au fur et à mesure du temps, à la suite du projet tout en étant guidée par l'équipe.

## 2) Le planning

Le planning a évolué depuis le départ car la suite de l'avancement devait dépendre des résultats trouvés au fil du temps. Les grandes lignes sont restées inchangées dans l'ensemble mais les résultats ont modifié la façon de réaliser le projet.

Plusieurs rapports intermédiaires ont dû être fournis et présentés à plusieurs occasions, notamment après la finalisation de chacune des grandes parties. Ils sont représentés par des jalons sur le planning GANTT ci-contre.

Au départ il était important de bien analyser les données pour ne rien laisser échapper et en parallèle réaliser des recherches sur des études similaires déjà réalisées. Par la suite, sélectionner et mettre au propre les données pour les utilisations futures. Puis les analyser pour en tirer des conclusions et finalement pouvoir récupérer un résultat prédictif selon une date donnée.

Voici le planning initial qui donne les grandes lignes du projet.



Figure 9 : Planning initial des tâches à réaliser

Ce planning initial, est un extrait du planning GANTT final, qui s'appuie sur celui-ci. Il reprend les grandes lignes de ce planning en le décomposant et l'ajustant selon les résultats obtenus au cours du projet.

Le planning GANTT définitif, avec chacune des tâches principales est présenté ci-dessous. Ce planning retrace les différentes étapes et leurs réalisations, ainsi que les dates de leur réalisation.

Tâches	Planning GANTT du Projet												
	Semaines 36-37 05-16 Sept	Semaines 38-39 19-30 Sept	Semaines 40-41 03-14 Oct	Semaines 42-43 17-28 Oct	Semaines 44-45 31-11 Nov	Semaines 46-47 14-25 Nov	Semaines 48-49 28-09 Dec	Semaines 50-51 12-23 Dec	Semaines 52-01 26-06 Jan	Semaines 02-03 09-20 Jan	Semaines 04-05 23-03 Feb	Semaines 06-07 06-17 Fev	Suite du projet
<b>1. Recherches, prises d'informations et analyses</b>													
1.1. Réaliser des recherches sur le sujet et se documenter													
1.2. Analyse des données d'entreprises													
1.3. Analyse des données de stationnement et compréhension													
1.4. Lecture d'articles de recherche sur le sujet													
1.5. Rédaction rapport d'analyse de données et présentation													
<i>J1 : Avancement des deux premiers mois</i>													
<b>2. Utilisation d'algorithme d'analyse des données</b>													
2.1. Définir les techniques de clustering utilisables													
2.2. Comparaison des éléments deux à deux													
2.3. Réalisation de l'algorithme 1 : Ascendant hierarchy													
2.4. Etude des jours types ressortants													
2.5. Réalisation de l'algorithme 2 : K-means													
2.6. Combinaison des deux algorithmes et jours types ressortants													
2.7. Rédaction rapport d'avancement (partie 2) et présentation													
<i>J2 : Avancement</i>													
<b>3. Résultats final et limites</b>													
3.1. Prise en compte des résidents													
3.2. Prise en compte de la fraude													
3.3. Résultats sous format json													
3.4. Résultats sous format visuels													
3.5. Rédaction du rapport de stage													
3.6. Ajout dans le JPS													
<i>J3 : Rapport de stage et soutenance</i>													

Figure 10 : Planning GANTT final du stage avec les dates de réalisation des principales tâches et les remises de rapports

On observe sur ce planning, que la première partie de recherche d'informations a dû être réitérée pour commencer la 3<sup>ème</sup> partie du stage. En effet, la deuxième phase de réalisation technique d'algorithmes pour organiser les jours en jours types a nécessité une phase de recherche sur le sujet et une bonne prise en main des données relatives aux visiteurs. La troisième phase de résultats a dû prendre en compte les données relatives aux résidents et aux fraudes. Ainsi, il a de nouveau fallu analyser les données pour en extraire les résultats attendus. De

plus, des recherches spécifiques et complémentaires concernant l'amélioration possible du taux d'occupation ont dû être réalisées.

Les différents jalons correspondent aux rendus intermédiaires et à leur présentation pour finaliser les trois grandes parties de ce projet. Les deux jalons de la partie 2 correspondent à la présentation des deux algorithmes établis sur le sujet.

Les étapes indispensables de ce stage sont multiples. Tout d'abord, l'analyse préalable des données doit être considérée comme indispensable. En effet, une mauvaise compréhension des données conduirait à des résultats biaisés. Il a été assez difficile de comprendre parfaitement les données au vu de mon inexpérience dans le domaine. Mais cette partie a été très constructive pour moi, notamment parce que les données ont dû être retravaillées et certains points avaient été mal compris au départ et avaient faussé mes résultats. Par exemple, les paiements en ligne sur un même arrondissement étaient arbitrairement regroupés sur la première zone de l'arrondissement ; il a fallu les redistribuer sur l'ensemble des zones correspondant à l'arrondissement. J'ai dû revenir sur celles-ci à posteriori, pour prendre en compte certaines subtilités. Cette difficulté est expliquée dans la partie finale de cette étude. Une autre étape indispensable est celle du choix de l'algorithme permettant la génération de jours types ainsi que son développement. Finalement, il a été décidé d'en développer deux, au vu des résultats du premier. Il a été beaucoup plus difficile de développer le premier algorithme que le second qui m'a semblé beaucoup plus facile à comprendre. La dernière grande étape consiste en l'ajout de données supplémentaires au taux d'occupation des visiteurs précédemment obtenu. Cette étape ne figurait pas dans le planning initial et a été ajouté au vu des résultats d'occupations alors incomplets. Ces données correspondent notamment aux données relatives aux résidents et à la fraude. La fraude a été très difficile à ajouter, comme on peut le voir dans le rendu de l'analyse. En effet, les données de fraude ne retracent pas exactement les réalités de terrain et il est très difficile de les ajouter aux autres. Pour terminer, l'étape finale correspond à la formalisation des résultats (définition d'un format de données en json pour l'exploitation future des résultats dans le calculateur d'itinéraires) ainsi qu'à l'application visuelle avec la création d'une carte temporelle de Paris (fonctionnalité qui pourrait se joindre à la coloration du trafic routier dans un contexte d'information voyageur).

### 3) Les contributions

Au départ, le projet avait été rapidement établi quelques années auparavant et avait été mis sur le côté dans l'attente d'une reprise. Il a donc fallu repartir du début, avec des données brutes et non traités de 2018 et un rapport datant de 2017. À la fin de ce stage, un moyen de mise en place de jours types a été établi, les différentes caractéristiques à prendre en compte pour considérer le taux d'occupation réel ont été identifiées : certaines comme la fraude ou la part de résidents ont été ajoutés, d'autres sont uniquement proposées pour un éventuel travail ultérieur. De plus, des données d'occupation et de temps de recherche sont retournés selon les données de 2018 à Paris. Ainsi, le projet du stage a abouti. Cependant, le projet peut évoluer dans le futur, avec l'arrivée de données supplémentaires si le projet est de nouveau proposé pour la ville de Paris ou bien en utilisant des données semblables sur d'autres villes. Evidemment, des données historisées récentes, voire l'instauration d'un flux de données temps réel, amélioreraient la précision des résultats et des estimations fournies. D'autres caractéristiques pourraient également être étudiées si disponibles, comme l'impact de la météo.

Ce projet était une étude annexe par rapport au travail de l'équipe TripPlanner et n'avait donc pas de lien direct avec les tâches de l'équipe. Ainsi, il pouvait être réalisé de manière autonome sans interférer avec les autres projets en cours. J'ai donc mené seul ce projet, tout en restant suivie par les membres de l'équipe qui m'aidaient à donner des directions à mes recherches. Les contributions exactes sont toutes celles établies dans le planning GANTT ci-dessus, ainsi que dans tout le rapport de stage.

La réalisation de ce stage peut être répertoriée comme une étude sur le fonctionnement d'obtention de jours types à partir de données datées, ainsi qu'une étude sur le taux d'occupation des parkings sur données réelles. Les résultats sont présents sous forme concrète avec des données d'occupations prédictives sur la ville de Paris ainsi qu'une visualisation sous forme de carte temporelle de temps de recherche selon les zones.

---

Mon travail peut être réutilisé directement avec les données json d'occupation selon les jours types définis dans le calculateur d'itinéraire. Il est donc nécessaire d'ajouter ces données au calculateur pour les rendre fonctionnelles dans celui-ci. De plus, le modèle prédictif peut être alimenté en amont avec les suggestions détaillées dans ce rapport, dans le cas où de nouvelles données seraient fournies.

#### 4) Les outils et technologies

Le planning s'organisait généralement par semaine, sans réelles dates butoirs, mais la suite dépendant des réalisations de la semaine précédente.

Le langage le plus approprié pour l'analyse de données est le python, car de nombreuses librairies sont utilisables pour traiter des données comme numpy, pandas, matplotlib, scikit learn ou encore scipy ou regex et plotly... Le logiciel principalement utilisé est Visual Studio Code qui permet de compiler le code et offre un débugger.

Un autre outil que j'ai utilisé est Git pour sauvegarder le code sur un serveur en plus de mon ordinateur et pour pouvoir consulter facilement l'historique de mon travail. Pour mieux utiliser Git, j'ai découvert Fork, logiciel qui offre une interface à Git et simplifie les actions de sauvegarde et de modifications du code et facilite la visualisation du travail. Les données sont en csv donc observables à partir de n'importe quel éditeur de texte, comme Notepad++. Les résultats ont été générés au format json, également observables avec un éditeur de texte. Pour les recherches réalisées, j'ai utilisé Google Scholar et Zotero afin de conserver les documents et les sources utilisées.

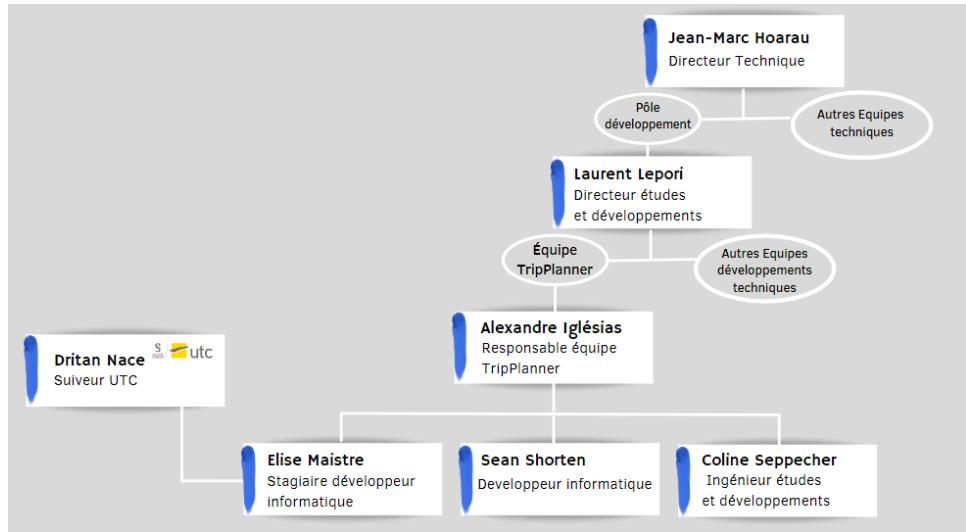
Pour réaliser mes différents rapports et afin de présenter mes résultats, j'ai utilisé Confluence disponible sur Atlassian proposé par l'entreprise. De plus, pour réaliser mon rapport de stage final, je me suis servie de Word provenant de Microsoft Office.

Afin de communiquer avec les autres, notamment lors du télétravail, j'utilisais Teams et la boîte mail Outlook.

Pour l'encadrement, deux méthodes différentes ont été utilisées pendant mon stage. Au départ, mon tuteur était en congé, l'équipe TripPlanner était alors encadrée par la responsable de l'équipe Data qui utilise une méthode extraite de la méthode agile. J'ai donc brièvement découvert cette méthode : chaque matin, chacun devait présenter ce qu'il avait fait le jour précédent et les éventuels points bloquants sur lesquels il avait besoin de communiquer. Cette réunion permettait de s'assurer que chacun savait ce qu'il faisait et de l'aider à avancer. Quand mon tuteur de stage est revenu, la stratégie TripPlanner a été mise en place, basée sur les motivations de chacun et sur un modèle de dynamique de groupe où chacun doit trouver sa place dans le projet selon ce qu'il juge être ses compétences et ses envies. Pour moi ce fonctionnement était plus difficile, ainsi on a décidé de mettre en place une réunion hebdomadaire d'une heure où je pouvais présenter mes résultats de la semaine sous forme d'un document ou de graphiques. A la suite de la réunion, on me faisait des retours et on discutait ensemble des différentes pistes de recherche à explorer ainsi que des attentes sur mes futurs résultats.

Pour vérifier et tester les résultats, j'ai plusieurs fois réalisé un extrait de données que j'utilisais afin d'augmenter la rapidité du processus. En effet, les données étaient lourdes et il était parfois nécessaire de vérifier des hypothèses sur un échantillon plus petit, par exemple sur deux semaines de données puis sur un mois.

On peut voir la place que j'ai occupée dans l'entreprise, ainsi que l'équipe dont j'ai fait partie et mes responsables à travers l'OBS suivant.



**Figure 11 : OBS présentant ma place au sein de Cityway durant ce stage**

## 5) Prise de recul

Le travail que j'ai réalisé durant ce stage est un travail à la fois de recherche et d'analyse sur des données réelles. Mon travail a permis de relancer le projet et pourrait, dans le cas où des clients tels que la ville de Paris ou d'autres collectivités seraient intéressées, mener à ajouter une nouvelle fonctionnalité au calculateur d'itinéraire. Dans le cas où la ville de Paris voudrait mettre en place la prise en compte du stationnement sur voirie, les résultats prédictifs d'occupation de la voirie par zones géographiques pourraient être réutilisés et potentiellement étoffés avec de nouvelles données. Ces nouvelles données, qu'elles aient pour provenance la ville de Paris ou une autre collectivité, pourront être étudiées de manière similaire à celle présentée dans cette étude. Ainsi, les recherches réalisées, les codes et les algorithmes établis pourront être réutilisés.

Les améliorations possibles sur ce projet sont multiples et dépendent tout d'abord, de l'ajout ou de la prise en compte d'autres données. En effet, les données s'étendent sur 11 mois, de janvier à novembre 2018. Ces données ne sont donc pas complètes sur une année calendaire et il serait intéressant d'obtenir les données concernant le mois de décembre 2018 afin de constituer une prédition plus fiable sur tous les jours de l'année. De plus, un moyen de vérifier la fiabilité des résultats serait d'obtenir une année de données supplémentaire pour comparer les prédictions à des mesures réelles. La vérification des prédictions est essentielle pour connaître la précision de l'analyse. En effet, dans le cadre de cette étude, des résultats ont été établis avec un inconnu important et il est très difficile de déterminer leur fiabilité. Avec une année de données pour comparaison, cette tâche deviendrait alors réalisable. Une autre manière de vérifier les données serait de conserver les données de paiements sur une semaine et en parallèle de vérifier par un comptage des véhicules le taux d'occupation réel à chaque instant, comme le propose certaines sociétés ou comme cela a été proposé dans certaines études. [8]

Il pourrait également être intéressant d'ajouter aux données actuelles d'autres caractéristiques. Par exemple, la prise en compte de données en temps réel, comme il était mentionné au départ, permettrait une nette amélioration des données, mais cet ajout semble peu probable dans un futur proche, au vu du coût de mise en œuvre. En effet, pour récupérer une occupation en temps réel des places, il serait nécessaire de positionner des capteurs sur chacune des places. Cette opération est très coûteuse, réalisée dans très peu de villes, souvent mise en œuvre sur une partie des places seulement et difficile à maintenir dans le temps. [22]

Par ailleurs, avec cette étude, les résultats ne sont extraits que des données et donc des plages payantes de 9h à 20h, du lundi au samedi. Il serait intéressant d'ajouter les périodes de nuit et de dimanche, ainsi que les jours fériés, par d'autres moyens, pour rendre les données prédictives plus intéressantes. On pourrait notamment s'appuyer l'étude de l'EGT [7] qui a réalisé des estimations de stationnement la nuit et les dimanches.

Comme aperçu dans cette étude, les données de fraude ne sont pas suffisamment détaillées, car elles ne sont estimées qu'à partir des données liées aux agents ayant attribués des amendes et non à la fraude réelle. De plus, ces

---

données sont décrites par arrondissement ce qui fait perdre en précision. Il serait intéressant d'estimer le taux de fraude réel par zone pour pouvoir le réutiliser dans les données. Ce taux pourrait être retrouvé à partir notamment du comptage de véhicules.

Les données de résidents pris en compte dans cette étude peuvent également être améliorées. Il serait par exemple possible de réaliser un sondage de la population parisienne comme cela a été fait dans l'EGT [7] et de poser la question du nombre de véhicules garés sur voirie selon les arrondissements. De plus, il serait utile de prendre en compte les résidents pendulaires et d'utiliser ces données. [10] La prise en compte du nouveau sondage de l'EGT 2022 en cours de réalisation pourrait apporter des informations intéressantes. [7]

Une autre amélioration possible serait de prendre en compte le prix du stationnement et d'ajouter un poids plus ou moins important dans le calculateur selon le montant, pour préciser les places les moins chères en priorité.

Il serait également intéressant d'apporter des informations directement au sein de l'algorithme permettant la production des jours types. En effet, les jours types ont été associés en fonction de caractéristiques liées à la période de l'année. Mais en ajoutant certaines caractéristiques comme la météo par exemple, il serait possible de trouver peut-être de nouveaux rapprochements, ou d'expliquer certains rapprochements. De plus, il pourrait aussi être utile de former des jours types pour les événements comme les manifestations (manifestations des gilets jaunes visibles sur les données de 2018) ou la coupe du monde ou encore les ponts, mais cette formalisation nécessiterait davantage de données pour être fiable.

La dernière amélioration possible serait d'ajouter les caractéristiques émises dans la troisième partie du rapport comme par exemple la prise en compte du taux de rotation, de la densité de places disponibles dans une zone, ou encore du remplissage des parkings en ouvrage à proximité.

Finalement, les données utilisées datent de 2018, elles viennent après la dépénalisation des FPS et après l'augmentation significative des prix du stationnement à Paris mais il serait intéressant de s'interroger sur la fiabilité de ces données étant données les politiques de changements de la mobilité à Paris effectuées ces dernières années, ainsi que l'impact du covid sur les comportements de déplacement avec par exemple l'arrivée du télétravail. De plus les prix ont réaugmenté ces dernières années.

L'objectif de cette étude est de simplifier la recherche des places de parking sur voirie pour l'utilisateur mais également de réduire la pollution en dissuadant l'utilisateur d'utiliser sa voiture ou en indiquant les endroits où il peut se garer plus rapidement, réduisant ainsi son temps de recherche d'une place libre. Si cette étude était ajoutée au calculateur d'itinéraire sur un territoire spécifique, elle permettrait une réduction de la pollution et orienterait les usagers dans la bonne direction.

# Etude du stationnement sur voirie

Le travail présenté a été divisé en plusieurs sous-tâches qui se suivent les unes les autres et se complètent, le tout en conservant un ordre précis lié à l'ordre chronologique. Il est intéressant de les comprendre et de suivre les liens qui existent entre ces parties.

Trois phases vont se succéder, une phase d'exploration, puis une phase de réalisation technique, pour finir avec une phase de réflexion sur les résultats obtenus.

Au départ deux tâches ont été réalisées en parallèle : la découverte et la compréhension des données d'une part, avec une phase de nettoyage des données aberrantes, et la recherche de documentation sur le sujet d'autre part. Ensuite, le sujet a été approfondi avec l'analyse d'articles de recherche sur le stationnement sur voirie.

La deuxième partie concernera quant à elle l'analyse plus détaillée des données pour aboutir à la génération de jours types particuliers. Cette section se concentrera sur les algorithmes étudiés et leur réalisation. C'est la partie la plus technique.

La dernière partie présentera les ajouts potentiels pour obtenir un résultat concluant. Cette partie expliquera également le format des résultats, ainsi que les limites à prendre en compte.

## Première partie : Recherches, prises d'informations et analyses

L'objectif de cette partie est de bien prendre en main le sujet pour se l'approprier et pour ne plus avoir de doute sur certains cas. Il est nécessaire de ne pas négliger cette partie car la compréhension des données est essentielle lorsqu'on réalise une étude d'analyse de données. Il faut donc tout d'abord être bien documenté.

## I) Recherches pratiques

La première étape consiste à bien maîtriser le sujet, bien comprendre le projet dans son ensemble et cela commence par prendre connaissance du fonctionnement de la ville en matière de stationnement. Et plus précisément comprendre les tarifs, les lois et le fonctionnement en vigueur pour l'année 2018, qui est l'année de référence de cette étude.

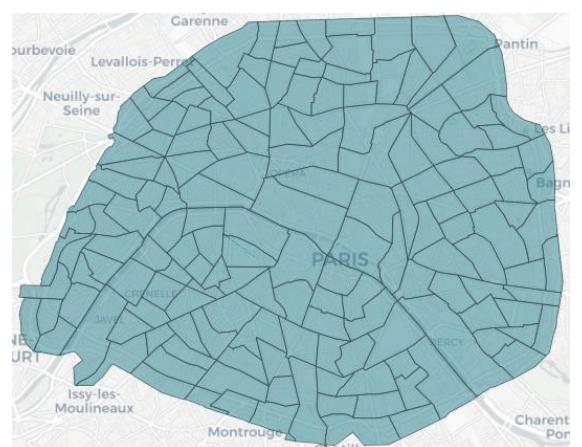
### 1) Stationnement sur voirie

Le stationnement sur voirie représente le stationnement des voitures en surface, aux places aux abords de la route prévues à cet effet. Nous nous intéresserons au stationnement payant, présent sur la voirie dans Paris.

La ville de Paris a été découpée en 160 zones de stationnement payants pour faciliter la gestion du stationnement résidentiel. Ces zones permettent à la ville de restreindre l'accès résidentiel au stationnement et aura l'avantage, dans le cadre de cette étude, d'avoir un meilleur aperçu des transactions de paiement. [24]

Il existe de nombreux types de places à Paris mais les données ne portent que sur celles payantes qui sont divisées en deux catégories :

- Les voies **mixtes** : elles sont destinées aux résidents et aux visiteurs, concernent la majorité des places et sont payantes de **09h00 à 20h00** toute la semaine et **gratuites les dimanches et jours fériés**.



*Figure 12 : Les 160 zones de stationnement à Paris*

- Pour les **résidents**, la durée de stationnement est limitée à **7 jours consécutifs**, au prix de 1€50 par jour. Le résident bénéficie de ce prix pour les **4 zones** autour du domicile et peut se déplacer en conservant son paiement sur l'une des 4 zones.

Pour des cas exceptionnels de **pollution** le stationnement résidentiel peut devenir gratuit sur décision municipale, lorsque le seuil d'alerte est dépassé. Néanmoins, les paiements sont quand même enregistrés.

- Pour les **visiteurs**, le temps de stationnement est limité à **6h00**. Les prix diffèrent selon les arrondissements et la durée (augmentation non linéaire). Les visiteurs peuvent également se déplacer en conservant leur paiement mais doivent rester sur la **même zone**.

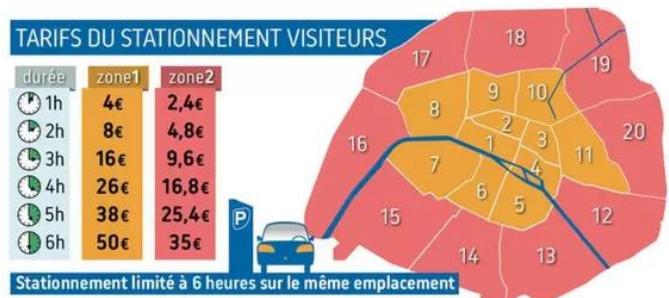


Figure 13 : Tarif du stationnement visiteur

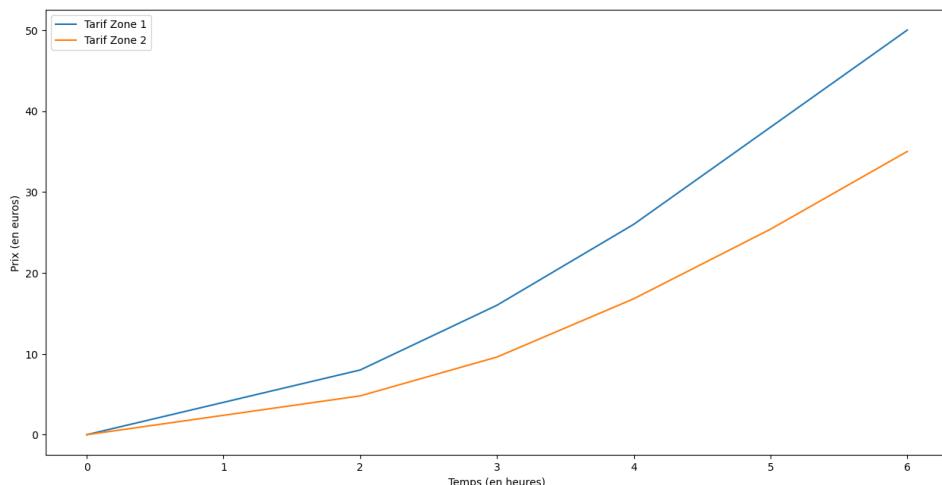


Figure 14 : Augmentation non linéaire du tarif visiteur en fonction des zones

- Les voies **rotatives** : elles sont réservées aux **visiteurs**, pour une durée limitée à **3h00** de stationnement et sont payantes de **09h00 à 19h00** et gratuites les dimanches et jours fériés. Le prix fonctionne comme pour les voies mixtes.

De plus, les stationnements de surface peuvent être pénalisés en cas d'infractions.

- 2) Amendes
- a) Le fonctionnement

La loi concernant les amendes en cas de non-respect du paiement au stationnement a été modifiée à partir du 1<sup>er</sup> janvier 2018. Ils sont maintenant nommés FPS pour Forfait Post-Stationnement et sont dépénalisés, c'est-à-dire que ce sont des entreprises qui s'en occupent et non plus les policiers. Dans la ville de Paris, ces FPS sont attribués par deux prestataires privés (Indigo et Urbis Park – Moovia et Streeteo) qui repèrent avec des voitures flasheuses les infractions et appellent ensuite des agents à pied pour venir, sur place, mettre l'amende. Le prix des FPS correspond à 6 heures de stationnement, c'est-à-dire à 50 euros au centre de la ville et à 35 euros en périphérie. Sur une seule journée, il est possible de recevoir jusqu'à 2 amendes. Dans le cas où le stationnement a déjà été payé plus tôt mais la durée est écoulée (cas de sous-paiement) le montant déjà payé est déduit du montant de l'amende et les 6 heures débutent à partir de l'heure d'arrivée réelle du conducteur.

---

## b) Les chiffres

Avec cette nouvelle réforme, une forte hausse des prix a eu lieu, l'amende précédente étant de 17 euros. Cette augmentation a certainement eu un impact sur la fraude dès 2018 ; que ce soit à propos de la mise en place ou de l'adaptation des automobilistes.

En effet, les sociétés étaient engagées à effectuer 75 000 contrôles par jour, mais d'après les chiffres seulement 10 500 FPS auraient été attribués par jour (3,2 millions dans l'année). Ces chiffres sont loin de ce qui était attendu et même inférieur à l'année précédente (4 millions pour l'année 2017). [30][30][30][30][30]

Deuxièmement, les automobilistes ont dû changer leur comportement en matière de paiement. D'après la mairie de Paris, le taux de paiement en 2018 serait passé de 9% à 20%. Ainsi, la fraude resterait tout de même très élevée.

Ces chiffres restent approximatifs et peuvent-être biaisés, il faudra donc vérifier ces données.

## 3) Résidents

A Paris, il existe un stationnement à tarif préférentiel pour les résidents. Cet abonnement permet aux habitants des arrondissements de Paris de stationner sur l'une des quatre zones autour de leur domicile. Cet abonnement est dématérialisé et contrôlé à partir de l'immatriculation du véhicule. Il est nécessaire que le domicile soit le domicile principal du demandeur et l'abonnement d'un résident est réservé à un seul véhicule.

## 4) Recherches d'articles scientifiques

Il est important de se documenter sur le sujet pour connaître les études déjà réalisées et pour pouvoir s'en inspirer et réutiliser certaines expériences. En effet, certaines études sur le même thème et avec le même type de données ont déjà été réalisées auparavant. Ces études ont eu lieu aux Etats-Unis avec des données en général moins importantes.

Des recherches concernant l'analyse de la fraude, la recherche d'une place de stationnement ou l'occupation de la voirie doivent également être documentées pour comprendre comment l'analyser et l'ajouter avec nos données. [15][19][22][23][28][3]

De très nombreux articles sont présents sur le sujet, pour la plupart des thèses réalisées aux Etats-Unis. Une grande majorité des articles n'ont pas les mêmes données pour tenter d'estimer le temps de recherche d'une place de parking sur voirie. Les données sont souvent en temps réel pour vérifier les résultats.

Plusieurs articles sur le sujet ont été réalisés par la ville de Paris ou par la région île de France à travers des enquêtes sur l'occupation et les activités, durant ces dernières années. [15][19][22][23][28]

Ces articles étudiés sont disponibles aux niveaux de la partie référence.

Après avoir étudié les documents et rapports sur le sujet, il est intéressant de réaliser une première étude statistique des données pour pouvoir par la suite les utiliser pour générer les résultats attendus.

## **II) Première analyse des données et statistique**

Les données fournies ont été récoltées auprès de la ville de Paris dans le cadre d'un ancien projet de Cityway, le projet Mobilité Intégrée en Ile de France (MI2). Ces données avaient pour vocation d'ajouter une nouvelle fonctionnalité dans le calculateur d'itinéraire : la prise en compte du temps de recherche de stationnement sur voirie. Des recherches avaient été menées puis rapidement abandonnées. Elles sont reprises dans le cadre de cette étude, qui a pour but d'évaluer ce temps grâce à la mise en place d'un modèle prédictif. Les données portent sur les transactions de paiement du stationnement et doivent permettre une estimation de l'occupation de la voirie en fonction de critères à déterminer.

La première étape consiste donc à analyser les données précisément pour comprendre comment les utiliser par la suite pour en tirer un résultat.

### a) Les données

Les données ont été récupérées sous la forme de 3 jeux de données au format csv. Ces données recensent la situation géographique des zones, les différentes emprises et les paiements effectués sur les 11 premiers mois de l'année 2018.

Le premier fichier reçu donne les périmètres géographiques des différentes zones de stationnements à Paris. Un périmètre est une ligne brisée exprimée par une suite de coordonnées (latitude, longitude). Ces géométries donnent une meilleure compréhension des données dans les arrondissements de Paris, et permettront de visualiser les résultats de l'analyse sur la carte. [24]

Le deuxième fichier nous indique plusieurs informations sur chacune des emprises de la ville. Une emprise sur la voirie correspond à un bloc d'emplacements d'une ou de plusieurs places.

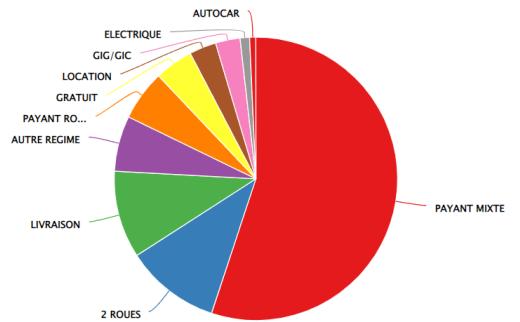
Les emprises sont répertoriées et indiquent le nombre de places de parking réelles qu'elles contiennent, les zones dans lesquelles elles se situent ainsi que la catégorie des usagers qui peuvent utiliser ces emplacements. Comme indiqué précédemment, l'étude portera sur la prise en compte des places de stationnement pour les voitures et payantes. Ainsi, les deux types de régimes conservés seront les places payantes mixtes et les places payantes rotatives. Comme illustré sur la figure ci-contre, la proportion de ces deux types de régimes (en rouge et orange) correspond à une majorité des places mais il reste beaucoup d'autres catégories d'usagers.

A partir de ces données d'emprises il est possible de visualiser le nombre de places mixtes et rotatives par zones, pour évaluer leur capacité et leur taille ainsi que de vérifier leur similarité.

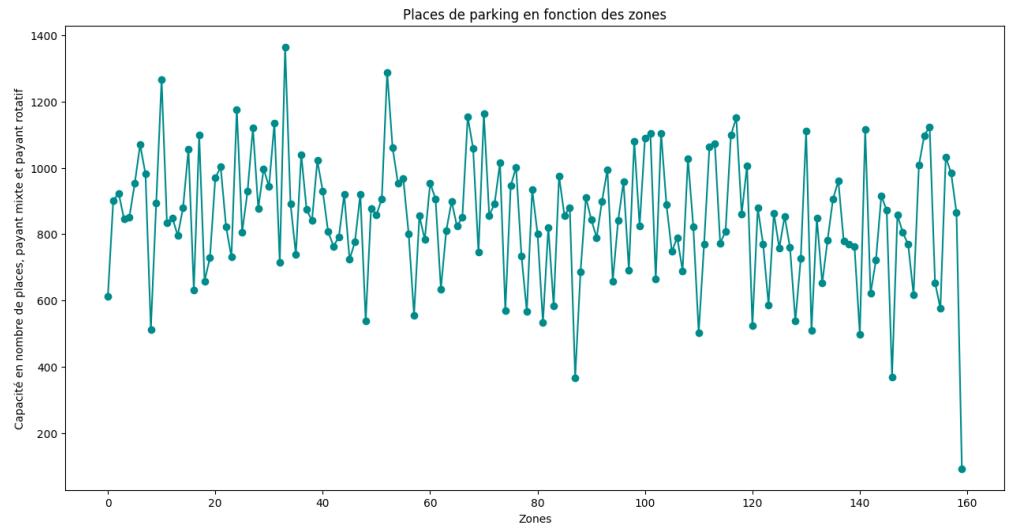
On remarque que les capacités ne sont pas identiques pour chacune des zones et qu'elles oscillent autour d'une moyenne de **850 places** par zone. On considèrera dans le cadre de cette étude que les zones sont relativement comparables entre elles. De plus, en faisant le décompte, on remarque qu'au total, il y a environ **136 000 places** payantes rotatives ou mixtes à Paris. C'est donc sur ces 136 000 places que va se poursuivre l'étude.

Ces deux premiers jeux de données sont disponibles en ligne sur le site d'opendata Paris et peuvent y être visualisés plus en détail. [14]

Le dernier jeu de données est composé de 33 millions de lignes sur les transactions effectuées sur les 11 premiers mois de l'année 2018 et n'est pas disponible en ligne. Il a été obtenu dans le cadre du projet de recherche MI2. Ces données sont composées des huit caractéristiques suivantes :



**Figure 15 : Camembert des différents régimes de places existants à Paris**



**Figure 16 : Capacité des zones de stationnement**

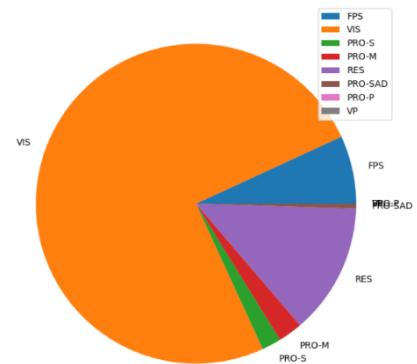
- Les temps de début et de fin de paiement avec une précision à la seconde. La date et l'heure de début sont enregistrées au moment de l'achat du ticket de stationnement, tandis que la date et l'heure de fin sont évaluées par rapport à la durée du paiement sur les plages horaires payantes.
- La zone dans laquelle le paiement a été effectué, ainsi que les zones de validité dans certains cas. La zone est composée de l'arrondissement dans lequel le paiement a été réalisé suivi d'une lettre ; le nombre de zones par arrondissement dépend de la taille de celui-ci et de sa capacité en nombre de places. Le découpage est visible plus haut.
- Le prix payé.
- Le profil de l'automobiliste, notamment visiteur, résident, professionnel ainsi que les FPS.
- Le fournisseur et le moyen de paiement utilisé (les horodateurs Parkeon disponibles sur la voirie, ou les différentes applications mobiles comme PayByPhone qui correspond à une grande partie des données...)
- Un identifiant unique pour chaque transaction

### b) Premières statistiques et analyses

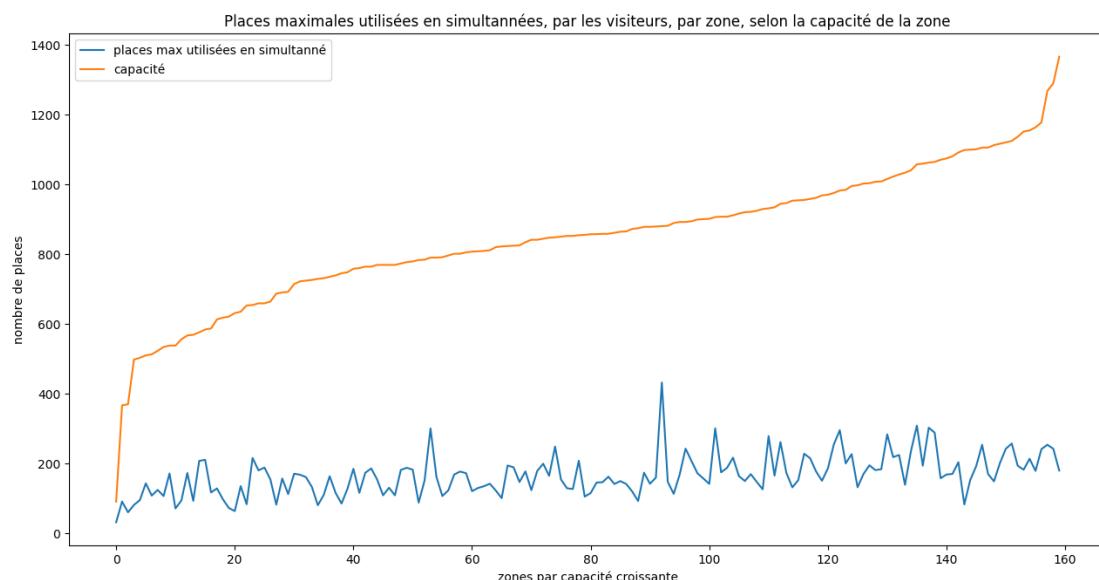
Ces données de transactions ont pu être visualisées et étudiées avant de les formater pour la formation de jours types à l'aide d'algorithme. Plusieurs études ont été menées pour comprendre les données.

Chacune des données ont été enregistrées selon un profil d'usager particulier et à chaque fois d'une manière spécifique. Il est intéressant d'observer le nombre de paiements effectués par profils, pour comprendre l'offre de ces données. Comme on peut le voir avec le camembert ci-joint, les visiteurs représentent la grande majorité des données, ensuite viens les résidents puis les FPS. Ces trois catégories seront celles étudiées dans cette étude.

Le nombre de places occupées a été décompté sur chacune des minutes de la journée. Ensuite, pour chacune des zones, le nombre maximal de places occupées en simultané, par les visiteurs, a été comparé à la capacité théorique de la zone, comme on le voit avec le graphique suivant. On remarque que le nombre de places occupées en simultané est largement inférieur à la capacité totale de chaque zone. Cette problématique représente la difficulté majeure de cette étude et sera expliquée et adressée dans la troisième partie.



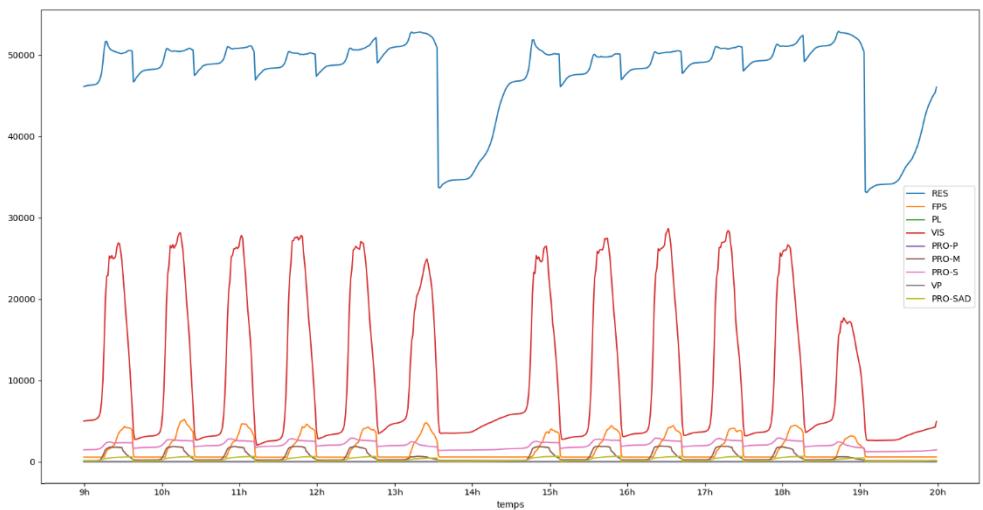
**Figure 17 : Proportion de chaque profil utilisateur dans les données reçues**



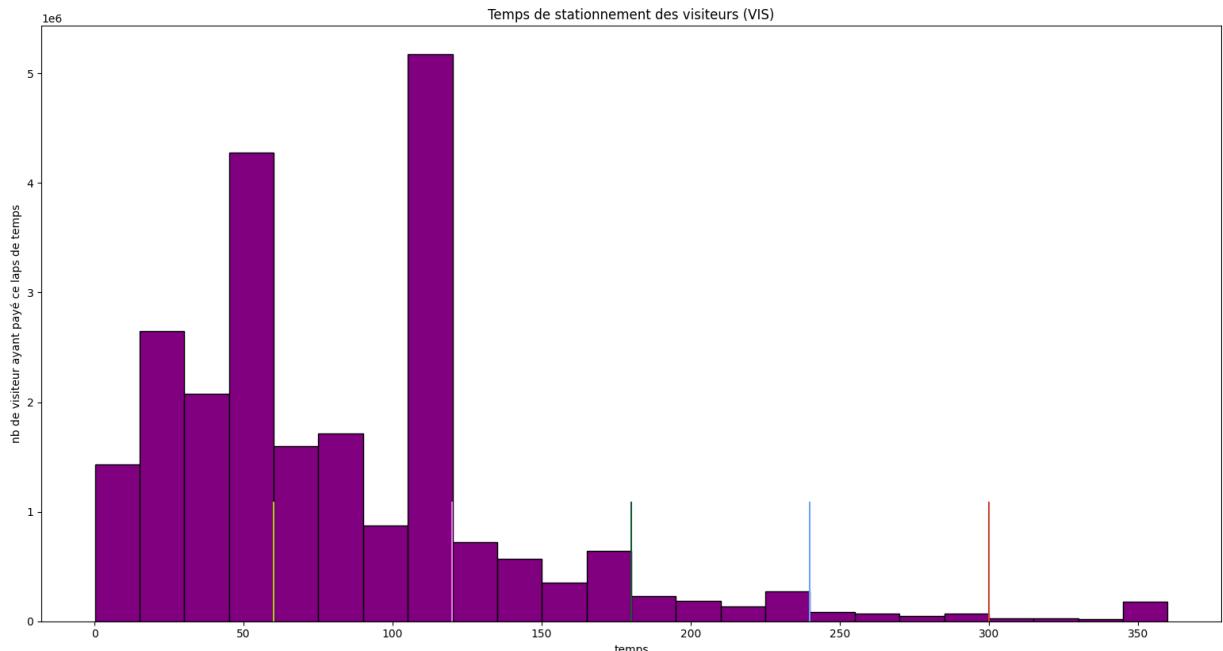
**Figure 18 : Maximum d'occupation des visiteurs en simultané sur les 11 mois de l'année pour chacune des zones, en comparaison avec la capacité de ces zones**

Il semble alors intéressant d'observer les différents taux d'occupation en fonction des profils des utilisateurs. On remarque que les résidents se retrouvent bien au-dessus des visiteurs : cela est dû au fait que les résidents sont comptabilisés pour une journée ou une semaine entière tandis que les visiteurs ne sont là que pour quelques heures ou minutes. Les FPS, en contrepartie, s'arrêtent brutalement à 20h.

Une dernière statistique intéressante est la durée de stationnement payée, par paliers de 15 minutes. On observe que les paiements de 2 heures et moins sont les plus fréquents ce qui semble le plus logique au vu des prix du stationnement sur voirie à Paris. Pour une grande partie des transactions, la durée de stationnement payée est d'exactement deux heures.



**Figure 19 : Courbes d'occupation simultanée, selon les types de profils, sur deux semaines**



**Figure 20 : Nombre de stationnements payés, par les visiteurs, par paliers de 15 minutes**

## Deuxième partie : Analyse technique des données et algorithmes

Après avoir réalisé une analyse approfondie des données et avoir effectué des recherches sur les sujets similaires, il devient important de cibler des jours représentatifs pour ensuite avoir une prédition efficace. En effet, si un jour appartient à un type particulier on peut faire l'hypothèse que le comportement d'un autre jour avec les mêmes caractéristiques sera similaire à celui-ci. Et à partir de ces résultats évaluer le temps de recherche pour en tirer un résultat qui doit être satisfaisant, malgré certaines limites.

### **III) Création des jours types**

L'objectif de cette partie est d'établir des **jours types** qui seront utilisés pour prédire les futures courbes d'occupation en fonction l'appartenance aux caractéristiques de tels ou tels jours types.

Il est intéressant d'établir dans un premier temps certains jours types à l'aide d'aprioris combinés avec les articles lus au préalable.

Nous établissons tout d'abord nos propres jours types par l'utilisation d'une méthode naïve.

#### a) Méthode naïve

Les observations personnelles laissent à penser que le jour de la semaine a un rapport avec la forme de la courbe d'occupation des places de parking. Cette idée est corrélée avec les articles qui créent des jours types comme les jours de la semaine du lundi au jeudi, les vendredis à part. [22] De plus dans d'autres articles et après analyse des courbes il s'avère que les samedis semblent très différents des autres jours. [7] Dans la littérature, les données s'étendent sur de courtes périodes et ne prennent pas en compte les vacances. Cependant, il semble intéressant de les considérer à part. De façon équivalente, les grandes vacances semblent être à part par rapport aux autres périodes de vacances.

On a donc 5 jours types : les petites vacances, les vacances d'août, et le reste divisé selon 3 groupes de jours de la semaine : du lundi au jeudi, les vendredis et les samedis.

Ces jours types naïfs permettent une première approche du problème et un moyen de comparaison avec d'autres résultats.

Il est maintenant temps de s'intéresser à des algorithmes plus précis, se basant sur les données reçues.

Pour trouver les caractéristiques liant différents jours entre eux, il est intéressant d'utiliser des algorithmes spécialisés. Pour cela, il faut tout d'abord posséder des données comparables entre elles. Il est donc nécessaire de nettoyer ces données de toute anomalies. Ensuite, l'objectif est de trouver des algorithmes permettant de les comparer les jours entre eux et de trouver des liens pour les rapprocher. Finalement, à partir de ces algorithmes, identifier les jours types convenables et réutilisable pour une prédition.

#### b) Formalisation des données d'entrées

Le premier objectif consiste à définir quelles sont les données importantes que nous voulons sauvegarder, et quelles sont les données qui pourraient permettre de trouver une corrélation entre certains jours. De plus, le format de ces données est primordial à définir. En effet, il joue un rôle dans la rapidité de lecture des données et de l'utilisation lors de l'exécution de l'algorithme. Par ailleurs, ces données doivent être nettoyées pour éliminer les valeurs aberrantes.

Tout d'abord, les différents profils observés dans l'analyse précédente ont été répartis en plusieurs fichiers distincts. Parmi ceux reçus, trois profils sont retenus : les visiteurs qui concernent la majorité des données et sur lesquels nous pouvons nous appuyer pour identifier des jours types, les résidents qui sont des données utilisables pour identifier le nombre d'abonnements résidents simultanés par zones et les FPS, ou amendes, qui

permettront une analyse plus détaillée de la fraude. Les données restantes concernant les différents types de professionnels ne seront pas utilisées pour cette étude étant donné leurs abonnements particuliers, utilisant des places particulières.

Les données concernant les visiteurs étant les plus détaillées et explicites, l'étude commencera avec l'utilisation de ces données, notamment lorsqu'il s'agit de définir des jours types. En effet, les données visiteurs sont plus représentatives des arrivées et des départs car les achats sont définis à la minute et non à la journée comme c'est le cas pour les résidents. Les données de résidents seront analysées ultérieurement.

- Formalisation des données

Les données sont réparties sur 11 mois de janvier à novembre. Les jours payants, donc pris en compte sont du lundi au samedi hors dimanche et jours fériés. Cela représente 276 jours de données qui peuvent être dénombrés par leur chiffre du jour de l'année de 1 à 333. Les heures payantes sont de 9h à 20h soit sur 660 minutes de la journée qui peuvent être ainsi positionnées entre la 540ème minute (9h) jusqu'à la 1200ème minute (20h) de chaque jour. Le temps de paiement des visiteurs doit être inférieur ou égal à 6h comptants (360 min). Ces 6h peuvent se retrouver sur une seule journée ou à cheval sur deux journées. En effet, lorsqu'un paiement est réalisé, il fonctionne sur les heures payantes ainsi le début et la fin d'un paiement peuvent être espacés d'une nuit, d'un dimanche, d'un jour férié ou d'une combinaison de ces cas. La fin du paiement ne peut se trouver que sur une plage horaire payante, lors d'un jour payant.

- Sélection des données utilisés

Les données conservées sont celles qui seront pertinentes pour la suite de l'étude. Les données dépendent fortement du temps ainsi les données temporelles sont essentielles. Le jour et l'heure de début ainsi que le jour et l'heure de fin sont nécessaires, la durée effectives (de maximum 6h) est également décomptée et conservée.

Le lieu permettant d'identifier plus précisément le paiement est également non négligeable, la zone est donc conservée. Les zones de validité qui précisent les lieux alentours où la voiture peut se situer, peuvent également servir pour la suite. Le type de paiement que ce soit sur horodateur (Parkeon) ou en utilisant une des applications comme PayByPhone peuvent être conservés.

Finalement le prix payé reste une donnée potentiellement intéressante.

- Nettoyage complet des données conservées

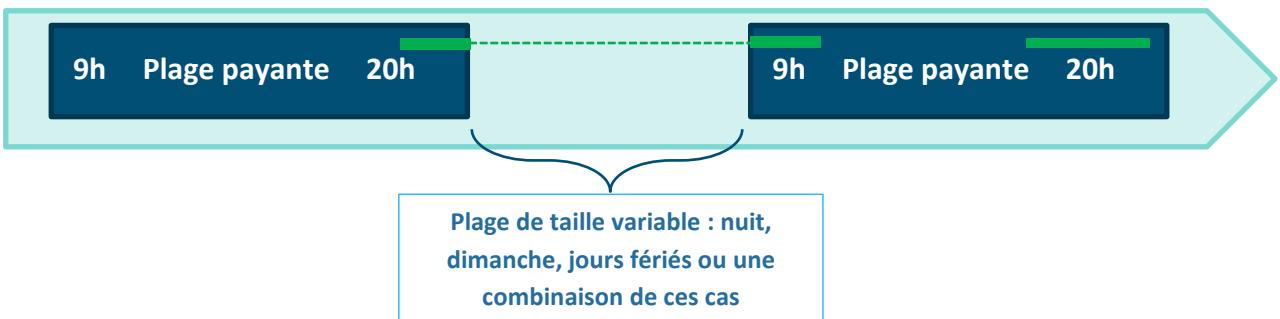
Les données choisies ne doivent comporter aucune erreur et être par la suite facilement et rapidement utilisables. Les plages payantes sont les jours qui ne sont ni fériés ni un dimanche et qui sont entre 9h et 20h sur ces jours-là. De plus, chacune des données doit appartenir à l'une des 160 zones.

Le début du paiement peut commencer en dehors d'une plage payante mais son décompte commence au début de la prochaine journée payante, au moment de la prochaine heure payante. Ainsi, il est plus intéressant d'enregistrer un début de paiement sur le prochain jour payant et au début du temps effectif soit à 9h, pour faciliter les futures utilisations de l'occupation réelle, étant donné que l'étude et les résultats ne porteront que sur les plages payantes des jours effectifs.

La fin du paiement ne peut être que sur les plages payantes, les données se terminant la nuit, un dimanche ou un jour férié sont des erreurs. De plus, la fin ne doit pas être avant le début de paiement, auquel cas c'est également une aberration.

Par ailleurs, la durée entre les deux jours ne doit pas dépasser six heures. Si ces deux jours se trouvent séparés entre deux plages non payantes il ne faut pas les supprimer, mais évaluer le nombre d'heure présentes sur la plage payante. La durée est également décomptée par palier de 15 minutes, d'autres périodes ne doivent pas être prises en compte.

## Visualisation du temps : jours payants ou gratuits :



**Figure 21 : Visualisation chronologique des plages payantes et gratuites**

Tous les cas étaient présents dans les données initiales, il était donc primordial de générer la prise en compte de ces erreurs. Sur les 25 millions de lignes récupérées concernant les visiteurs, 100 000 ont pu être retirées car correspondant à des erreurs. Ce taux d'erreur peut s'expliquer par la gestion de l'enregistrement des données et également par la diversité des machines de paiement et des erreurs potentielles dans l'enregistrement des paiements.

- Définition du format de données

Les données ont été récupérées à partir d'un fichier de données csv de 33 millions de lignes et retournées sous un autre fichier csv avec un nouveau format de données.

Ce nouveau format de données a été réalisé pour améliorer la rapidité et diminuer l'espace mémoire utilisé. Après plusieurs tentatives, il s'est avéré que le format de données utilisant la date et l'heure en string, n'était pas adapté, prenant un temps considérable et trop d'espace mémoire. Afin d'optimiser le processus, le format des données a été modifiée sous forme d'entiers, la date du jour de l'année, l'heure de la journée et la durée en minutes. Les dates et heures ont été ajustées et modifiées pour correspondre aux réelles plages payantes ainsi que la durée ne dépassant ainsi jamais 360 minutes.

```
enddate = datetime.datetime.strptime(line[1], '%Y-%m-%d %H:%M:%S')
date_fin = enddate.date().timetuple().tm_yday
emonth=int(line[1][5:7])
eday= int(line[1][8:10])
tab_jour_annee = [0,31,59,90,120,151,181,212,243,273,304,334]
date_fin = eday + tab_jour_annee[emonth-1]
```

Première tentative

Deuxième tentative

**Figure 17 : Explication de deux manières de lire une date**

La première tentative prenait un temps considérable, car les données conservaient leur format de date et pour les lire il fallait les convertir de string en date et les utiliser avec des fonctions de dates et de même les enregistrer par le format de date. La deuxième tentative, lire directement caractères par caractères dans le fichier csv et déduire le jour de l'année à partir d'un tableau d'entier donné au préalable indiquant le nombre de jours par mois. Il faut juste faire attention au cas des années bissextiles.

- Le cas des résidents

Pour les résidents, la méthode employée ressemble à celle des visiteurs mais se différencie sous certains aspects. En effet, les résidents payent pour une à sept journées consécutives. Deux formats de données différents sont possibles selon les providers. Les jours peuvent être payés comme une journée entière de 9h à 20h. Ou bien, les jours sont payés d'une heure de la plage payante jusqu'au lendemain à la même heure. De même que pour les visiteurs, la date de fin devait se trouver sur une plage horaire valide.

Les deux changements d'heure de l'année 2018 avaient également perturbé l'horodateur et des erreurs s'y étaient glissées. Une nouvelle information a également été ajoutée pour les résidents, ce sont les jours pollués, qui sont des jours où le stationnement est gratuit, uniquement pour les résidents pour inciter l'utilisation des transports en communs.

La méthode employée a donc été de compter le nombre de jours gratuits dans l'intervalle et de définir en fonction du format le nombre de jours payés pour vérifier que la durée ne dépasse pas les 6 jours payants.

- Difficultés rencontrées

Plusieurs difficultés ont été rencontrées. Tout d'abord, les cas valides et invalides ont dû être établis en amont et certains cas ont été trouvés au cours de l'analyse en utilisant des cas de test. Les données non valides sont quand même enregistrées dans deux autres fichiers annexes, en distinguant les cas où le temps de paiement effectif est supérieur à 6 heures et les cas où les valeurs sont aberrantes.

Une autre difficulté a été rencontrée lors de la séparation des jours entre eux pour définir la taille des trous entre chacun des jours et réussir à le coder sans se perdre en pensant à chacun des cas possibles.

Une autre difficulté a été lié à la première utilisation d'un dataframe pour traiter les données ce qui été à la fois beaucoup trop long au vu de la quantité de données (plusieurs heures) et prenait trop d'espace mémoire (beaucoup de crash au cours du processus). Il a fallu recommencer entièrement tout le code pour traiter les données avec l'aide d'une fonction read csv qui lit le fichier csv ligne par ligne.

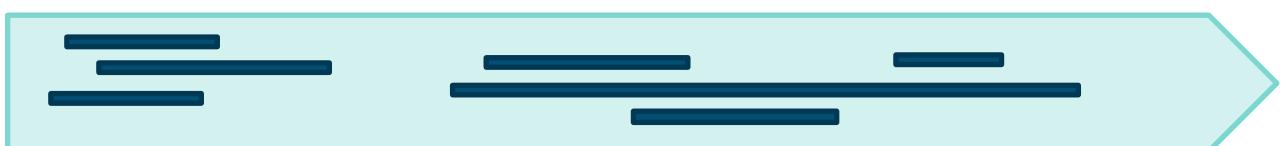
Le cas des formats des dates a également posé problème. La lecture directe d'une date au format de date prenait un temps considérable et baissait fortement les capacités de l'algorithme, il a fallu trouver une autre solution.

Pour les résidents, la difficulté a été la différence avec les données visiteurs, notamment au niveau du décompte du temps de stationnement. De plus, l'erreur liée au changement d'heure n'a été découverte qu'après, lors de l'analyse des fichiers d'erreurs. De même, pour les jours pollués, l'information les concernant n'a été découverte qu'ultérieurement lors de l'analyse des données et au vu d'un pic important des recherches ont été menées. L'accès à ces jours a été difficile car non répertoriés par la ville de Paris.

Après avoir formalisé, nettoyé et conservé ces données, il est nécessaire de les visualiser et de comprendre leur occupation.

c) Courbe du nombre de paiements visiteurs sur la journée

Pour utiliser les données, il est intéressant de les visualiser sous forme de courbe. Pour cela, pour chaque jour, pour chaque minute de la journée, un grand dictionnaire s'incrémenté de 1 afin d'enregistrer en simultané, toutes les personnes présentes sur la journée



*Figure 18 : Figure illustrant le décompte des places occupées en simultané, dans le temps*

Un dictionnaire composé de 276 jours payants comme clés et d'un tableau de 660 minutes comme valeurs, ces 660 minutes correspondent aux minutes de la plage payante, de 9h à 20h. Ce tableau est utilisé pour conserver les données et y avoir accès plus facilement. Ainsi, si une voiture visiteur a réalisé un paiement durant un laps de temps sur une journée donnée, ce laps de temps est incrémenté. Il est alors possible de représenter ces courbes qui correspondent aux visiteurs ayant réalisé un paiement en simultané pour chacun des jours, comme on peut le voir si dessous.

Grâce au formatage réalisé à l'étape précédente, il est ainsi plus rapide et plus facile d'enregistrer les données. Il est tout de même nécessaire de décomposer les lignes en deux types selon que le paiement s'étale sur plusieurs jours ou sur le même jour pour incrémenter précisément les minutes occupées.

On observe sur cette courbe que certains jours réagissent de manière différente mais qu'une tendance semble se démarquer. On remarque également que les paliers de 15 minutes payés par les utilisateurs se démarquent très nettement au début de la journée car tous les paiements débutent à 9 heures, il faudrait alors lisser ces valeurs pour avoir une vision plus réaliste des taux réels d'occupation et s'éloigner du modèle des paiements. Le lissage effectué est donc réalisé sur 15 minutes en utilisant une moyenne. Pour les 15 premières minutes le résultat est fixe

car le lissage nécessite les premières valeurs pour déterminer les suivantes, c'est donc une moyenne sur ces 15 minutes.

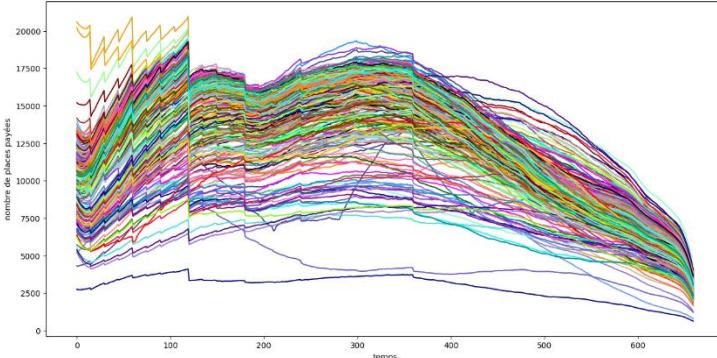


Figure 24 : Courbe d'occupation des visiteurs par jours

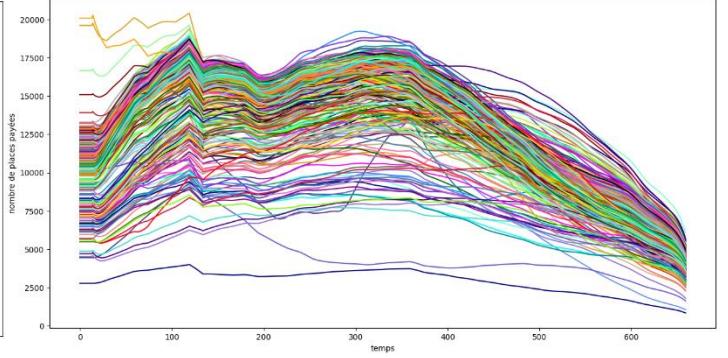


Figure 19 : Courbe des visiteurs simultanés par jours lissés sur 15 minutes

Le lissage permet alors d'obtenir des courbes plus réalistes. Les données sont enregistrées de la même manière que celle non lissées par minutes. Elles sont alors prêtes pour être comparées dans les algorithmes.

#### d) Choix de la distance de comparaison

L'objectif est de réutiliser ces données de présence au cours de la journée sur chacun des jours pour établir des jours similaires, appelés jours types. Ces caractéristiques doivent être facilement reconnaissables sur d'autres jours pour utiliser ces jours types comme jours prédictifs. Le plus simple est donc de commencer par considérer les jours comme une date et d'analyser leurs liens entre elles à partir de notre calendrier.

Un algorithme de clustering est alors celui qui correspondrait le mieux à la problématique. Pour réaliser ce type d'algorithme qui regrouperait les jours en fonction de leur similarité, il est nécessaire de pouvoir comparer les jours entre eux. Il existe plusieurs moyens de les comparer mais pour cela, il faut établir une fonction de distance. Cette fonction de distance compare tous les éléments deux à deux et permet d'établir un moyen de comparaison entre les données. Pour enregistrer la distance de dissimilarité, c'est-à-dire l'éloignement des éléments deux à deux, il est nécessaire de connaître le degré de similarité entre deux classes. Pour cela, on utilise généralement une fonction de distance.

Pour rappel, une distance est définie par trois lois fondamentales qui sont la symétrie, la séparation et l'inégalité triangulaire. Les distances les plus connues sont la distance de Manhattan, la distance euclidienne, la distance maximale ou encore de Canberra ou de Minkowski.

Les distances les plus utilisées dans le domaine du clustering sont la distance de Manhattan et la distance euclidienne.

Distance euclidienne [29]

$$d(X_{i_1}, X_{i_2}) = \sqrt{\sum_{j=1}^P (X_{i_1}^j - X_{i_2}^j)^2}$$

Distance de Manhattan [5]

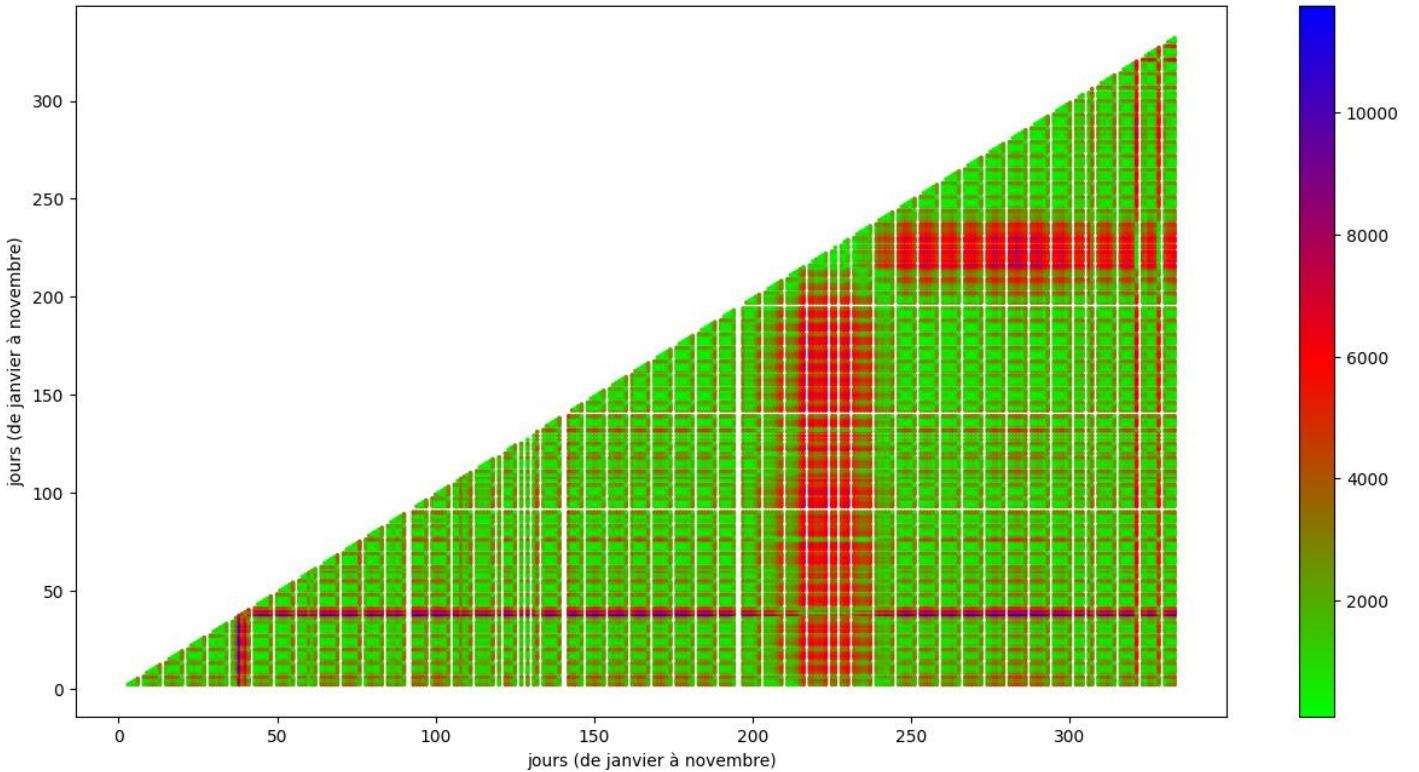
$$d(X_{i_1}, X_{i_2}) = \sum_{j=1}^P |X_{i_1}^j - X_{i_2}^j|$$

L'espace possède 660 dimensions car la comparaison est effectuée sur les minutes entre chacun des jours deux à deux. Lorsque la dimension est élevée (supérieure à 20), il est intéressant d'utiliser la distance de Manhattan plutôt que la distance euclidienne, la classification étant ainsi meilleure et plus proche. De plus, malgré que les minutes aient été lissées, il se peut que certaines variations soient plus importantes car les données ne sont pas toujours parfaites et comportent certaines irrégularités qui ne sont pas nécessairement des irrégularités de terrains. Il n'est donc pas nécessaire de mettre l'accent sur ces variations ou de les augmenter artificiellement, ce que fait la distance euclidienne contrairement à la distance de Manhattan.

La distance choisie pour effectuer le clustering est donc la distance de Manhattan.

Chacune des 660 minutes d'un jour sont comparées avec les 660 minutes de l'autre jour et les différences absolues entre chacune des minutes effectives de la journée sont additionnées, pour ainsi avoir la distance globale entre les deux jours. Cette opération est réalisée pour tous les jours deux à deux. Cette méthode permet une bonne estimation de la distance en prenant en compte l'occupation de chacune des minutes de la journée.

On peut observer le résultat à partir du graphe suivant, qui met en avant les différences grâce à une échelle de couleur. Plus les distances sont faibles, plus les jours sont proches, plus la couleur est verte. Lorsque les jours sont rouges ou bleus la distance entre les jours est très élevée (plus de dissimilarité). On peut déjà identifier quelques jours qui se démarquent plus que les autres.



**Figure 26 : Comparaison des jours deux à deux par distance de Manhattan, les 333 jours sont affichés dans l'ordre chronologique**

Les traits blancs correspondent aux dimanches et jours fériés, qui ne contiennent aucune donnée. D'après cet affichage, on observe les premiers jours types apparaître, comme les samedis qui se démarquent clairement. On voit également les événements forts de l'année 2018, comme les fortes neiges à Paris qui tirent vers le bleu en bas (6 au 10 février), les manifestations de gilets jaunes à droite (samedis 17 et 24 novembre), ou encore les grandes vacances qui se démarquent visiblement.

Cette première analyse rapide a permis d'identifier des jours particuliers grossièrement. En utilisant un algorithme de clustering, il serait peut-être plus évident de classifier certains jours ensemble plutôt que d'autres et d'établir une meilleure répartition.

#### e) Algorithme de classification hiérarchique ascendante

Après avoir établi un moyen de comparaison entre chacun des éléments (ici des 276 jours effectifs), il faut maintenant tenter de regrouper ces données en fonction de ces distances. Pour définir quelles distances seraient les plus intéressantes à associer, il est judicieux de se tourner vers de l'apprentissage non supervisé, en utilisant une méthode de clustering. Cette méthode consiste en un regroupement de données (ici des courbes de 276 jours) en différentes catégories ou classes proches par l'utilisation d'un algorithme.

Il existe deux méthodes de clustering assez connues, l'algorithme de classification hiérarchique ascendante (ou descendante) et celui des k-moyennes (ou k-means). [22][18]

Les deux algorithmes sont plutôt proches mais divergent sur certains points. La différence entre les deux algorithmes repose principalement sur une connaissance de départ. K-means nécessite de définir à l'avance le nombre de classe, ou ici de jours types, ce qui paraît difficile au départ.

Il semble donc intéressant de commencer avec l'algorithme de hiérarchie ascendante. En effet, de prime abord, il est difficile de connaître le nombre des classes voulues (jours proches, aussi nommés clusters) d'autant

plus avec un aussi grand nombre de données. En contrepartie, cet algorithme prend plus de temps et plus d'espace mémoire, au vu des nouveaux calculs effectués pour chaque nouvel ajout dans un cluster.

Le principe de hiérarchie ascendante est qu'elle est ascendante contrairement à la hiérarchie descendante, elle part du principe que chaque élément au départ représente une classe, un cluster séparé. Au fur et à mesure l'algorithme va regrouper les classes de distance les plus faible entre elles. A la fin du processus, toutes les classes sont regroupées entre elles, ainsi la définition du nombre de cluster final (ou jours types définitifs) vient après et peut-être choisi en fonction de la distance maximale voulue au sein d'une classe. [16]

### ALGORITHME DE CAH : [22][26]

*Début :* Tout d'abord considérer chacun des éléments comme distincts les uns des autres, au départ pour n éléments il y a n classes.

1. Trouver la dissimilarité pour chacun des éléments deux à deux à l'aide de la fonction de distance (cf distance de Manhattan)
2. Regrouper les deux éléments ayant le moins de dissimilarité entre eux selon la fonction de regroupement choisie (cf voir ci-dessous : fonctions de regroupement)
3. Récursivement : Tant que tous les éléments ne forment pas une classe unique :
  - a. Calculer la matrice des distances entre les classes en respectant la fonction de distance ainsi que la fonction de regroupement définie au départ
  - b. Fusionner les deux éléments les plus proches (ayant le moins de distance) selon la méthode de regroupement

*Fin :* Finalement tous les éléments sont regroupés en une seule classe avec des distances de regroupement en son sein différentes qui permettra la construction de l'arbre

Après la réalisation de l'algorithme, il ne reste plus qu'à définir la distance maximale voulue entre les éléments d'une même classe pour trouver le nombre de jours types.

Dans la littérature, cet algorithme est présenté avec une dimension. Dans notre cas, les données possèdent 660 dimensions, il a donc fallu adapter l'algorithme aux données ce qui fut un travail fastidieux et difficile au vu de mon inexpérience au départ.

Dans l'exemple, les données sont représentées comme suit.

$$X = \begin{pmatrix} \text{1ère min} & & & \text{660ème min} \\ (9h00) & & & (20h00) \\ x_{(1,1)} & x_{(1,2)} & x_{(1,\dots)} & x_{(1,n)} \\ x_{(2,1)} & x_{(2,2)} & x_{(2,\dots)} & x_{(2,n)} \\ \dots & \dots & \dots & \dots \\ x_{(m,1)} & x_{(m,2)} & x_{(m,\dots)} & x_{(m,n)} \end{pmatrix} \rightarrow \begin{array}{l} \text{jour 1} \\ \text{jour 2} \\ \vdots \\ \text{jour m} \end{array}$$

Figure 28 : Format des données de clustering

Pour regrouper les éléments entre eux, il est nécessaire d'avoir une méthode de regroupement. Il en existe plusieurs qui ont été testées sur les données. Après avoir trouvé la plus petite distance (le moins de dissimilarité) entre tous les clusters deux à deux, ces deux anciens clusters vont se regrouper pour n'en former qu'un seul. De plus, l'étape suivante est de nouveau un calcul des distances deux à deux. La manière de regrouper les clusters peut être différente. En voici quelques exemples : [16]

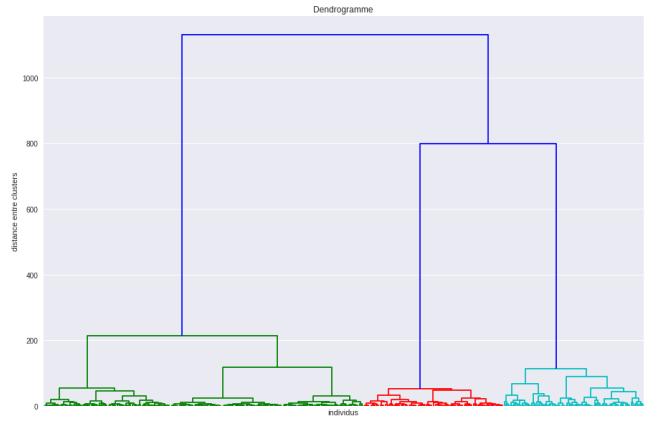


Figure 27 : Représentation de l'algorithme sous forme d'un dendrogramme

- *Distance des barycentres* : Le regroupement effectué est une moyenne pondérée en fonction de la taille de chacun des clusters. Les deux clusters ont une taille dépendant du nombre d'éléments qui les composent. Cette taille pondère la moyenne. Cette pondération est effectuée sur chacune des 660 minutes prise une à une. Ce nouveau calcul correspond à une nouvelle ligne ajoutée à la matrice, tandis que les deux anciens clusters (deux autres lignes de la matrice) sont supprimés de la matrice. Ainsi, la taille de la matrice diminue d'un élément à chaque itération. A la suite de la formation de cette ligne, qui correspond donc au barycentre entre ces éléments, les distances de Manhattan peuvent de nouveau être calculées. Ces distances sont donc calculées entre chacun des barycentres des clusters.

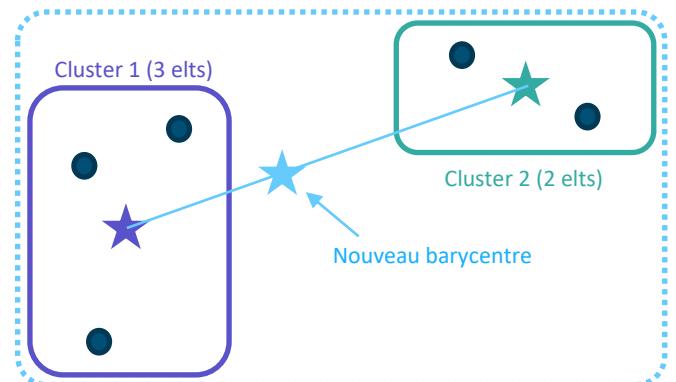


Figure 29 : Explication de la méthode de regroupement par distance des barycentres

- *Distance moyenne ou lien moyen* : La distance moyenne se différencie de la distance des barycentres par le fait que le centre du cluster n'est pas recalculé à chaque ajout sur les 660 minutes mais est fonction des comparaisons déjà réalisées. Ainsi, son temps de calcul est amoindri. Ces sont les distances deux à deux qui sont conservées et modifiés. Les deux clusters ayant la distance la plus faible sont considérés comme un nouveau cluster à part entière. La distance de ce nouveau cluster avec les autres clusters est calculée comme une moyenne pondérée dépendant de la taille de chacun des deux anciens clusters précédent et de leur distance respective avec chacun des éléments. Puis, toutes les comparaisons individuelles des parties de ce cluster sont supprimées pour ne garder que les comparaisons de ce nouveau cluster avec les autres et les autres comparaisons de cluster entre eux, pour réitérer le processus.

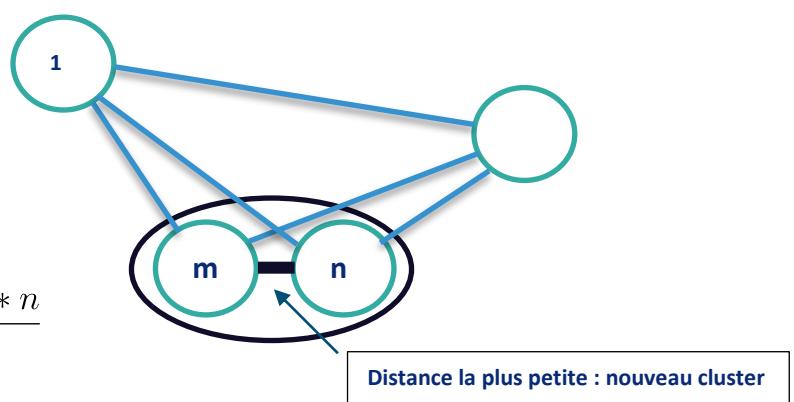
La distance la plus petite entre ces 4 clusters, correspond à celle entre les clusters de taille  $m$  et  $n$ . On créer alors un nouveau cluster de taille  $m+n$ . La distance de ce nouveau cluster avec  $C_1$  est calculée à partir des anciennes distances entre  $C_1$  et le cluster de taille  $m$  ( $d_1$ ) et  $C_2$  et le cluster de taille  $n$  ( $d_2$ ). Ainsi, la distance entre  $C_1$  et le nouveau cluster correspond à la formule suivante :

$$\frac{d_1 * m + d_2 * n}{n + m}$$

Ensuite les distances  $d_1$  et  $d_2$  sont supprimées et seule la nouvelle distance est conservée.

- *Distance minimale (maximale)* : La distance minimale (et maximale) ressemble beaucoup à la distance moyenne ci-dessus. Seulement dans ce cas, la distance conservée entre le nouveau cluster et  $C_1$  est la distance la plus petite (la plus grande) parmi  $d_1$  et  $d_2$  et non plus la moyenne pondérée.

Les algorithmes ont été réalisés de manière récursive jusqu'à ce qu'il n'y ait plus de comparaison à effectuer. Il s'est avéré, après analyse des données que la distance moyenne était la plus révélatrice des jours types. Pour mieux comprendre les résultats, il est intéressant de les observer à partir de dendrogrammes et de leurs courbes associées



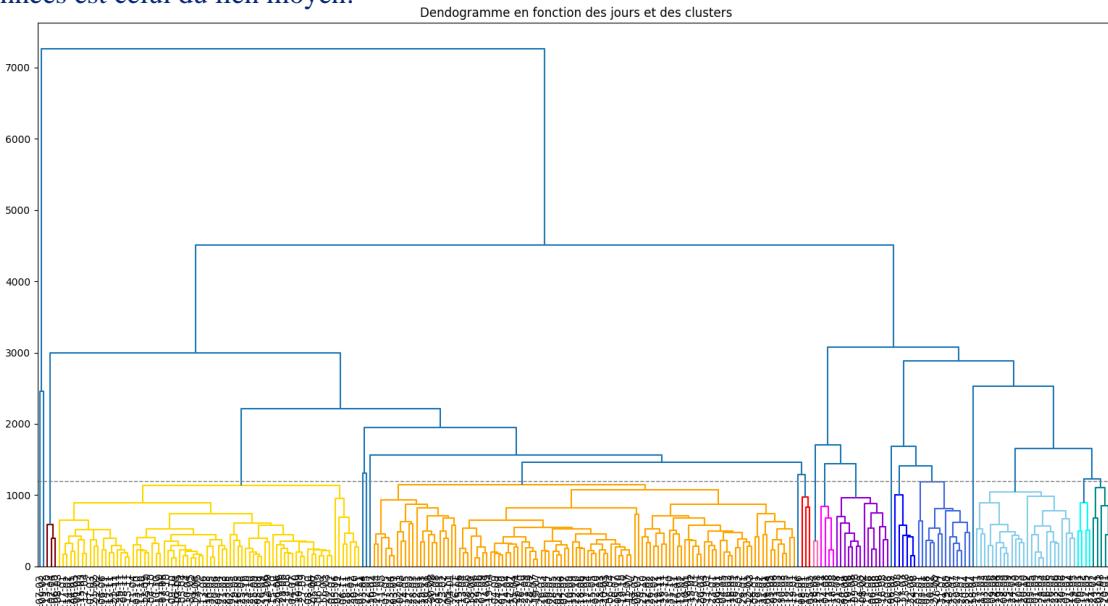
- Difficultés

Plusieurs difficultés se sont posées lors de la réalisation de ces algorithmes. En effet, la compréhension de ceux-ci s'est révélée difficile, d'autant plus qu'il était nécessaire d'adapter les algorithmes, prévus pour des données de taille 1, aux données réelles de taille 660. Une difficulté supplémentaire réside également dans l'explication de ces algorithmes pour vérifier la bonne compréhension de ceux-ci. La réalisation de ceux-ci m'a pris plus de temps que la durée estimée au départ, car j'ai dû découvrir ces algorithmes et les recoder entièrement.

#### f) Dendrogrammes et courbes associées

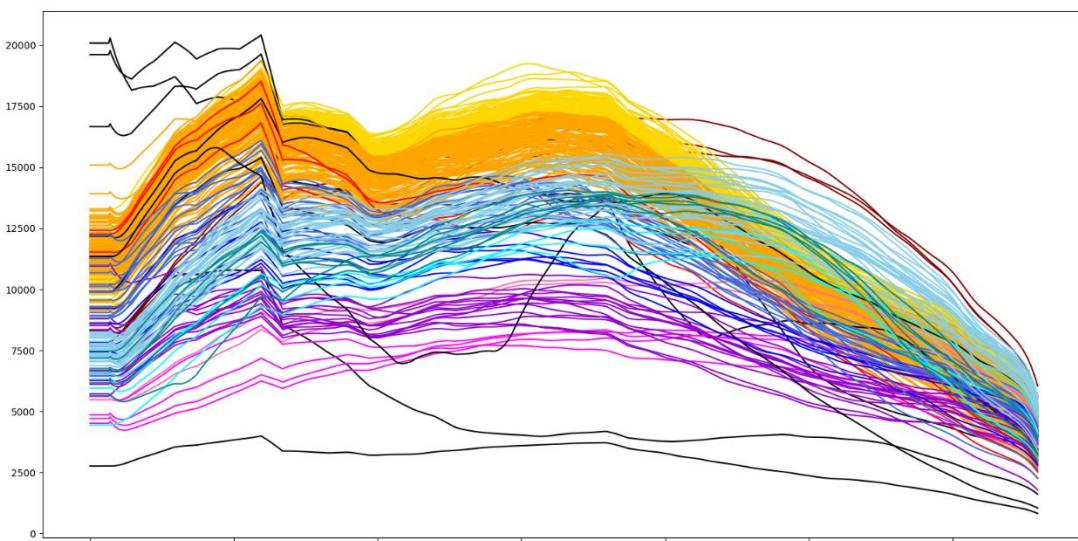
Un dendrogramme est une visualisation, sous forme graphique des différents clusters formés et de la distance entre les éléments et les clusters. Un dendrogramme permet d'afficher toutes les étapes de l'algorithme et de les regrouper dans un arbre pour permettre une meilleure analyse et de décider visuellement des différents jours types. Il est important de définir par observation la distance maximale voulue entre les éléments au sein d'un même cluster. Elle définira le nombre de cluster final.

En analysant les différents dendrogrammes et les courbes associées, il transparaît que le meilleur algorithme avec ces données est celui du lien moyen.



*Figure 30 : Dendrogramme issu de l'algorithme d'ascendance hiérarchique avec lien moyen*

Ce dendrogramme permet de visualiser certains jours types qui pourront être utilisés dans le choix final. Il est nécessaire d'analyser en détail chacun des clusters formés, pour voir quel lien peut en être tiré.



*Figure 31 : Courbes avec couleurs associées aux dendrogrammes ci-dessus*

Les courbes noires sont des clusters à part qui ont trait à un seul jour particulier, leur distance avec les autres éléments ou clusters étant trop importante. Ces jours sont en réalité des jours particuliers comme certains jours de neiges, des jours durant certains ponts ou encore des bugs de données sur un jour particulier notamment. D'autres jours particuliers comme les périodes de gilets jaunes ou certains autres ponts sont mélangés à certains clusters, n'ayant pas de lien avec eux. Les clusters se distinguent par le jour de la semaine, la période de l'année ou encore la présence de vacances.

En analysant plus précisément ces données tout en conservant un moyen d'affichage clair indiquant le jour de la semaine par une lettre, la date précise avec son mois et la présence de vacances symbolisée par le signe \*. Les données sont alors décortiquées pour être par la suite réutilisées.

La technique : le nombre de chaque jour de la semaine est représenté sous un tableau, de même pour les mois de l'année, le taux de vacance est retenu ainsi que la variabilité intra-classes. Ces données sont associées à la courbe et visualisées avec la courbe

#### Analyse :

On observe que les vacances ne se distinguent pas des autres jours, hormis les vacances d'août contrairement à ce qu'on aurait pu penser (cf méthode naïve). De plus, les samedis sont distincts des autres jours de la semaine. Les jours de la semaine sont eux-mêmes découplés particulièrement. En effet, les lundis et vendredis se regroupent ainsi que les mercredis et jeudis, les mardis sont présents dans chacun des deux groupes (sur le dendrogramme, groupe gold et orange). Il est intéressant d'analyser ce cas pour mieux trier ces jours et notamment savoir où situer les mardis. Comme on les voit sur la courbe ci-contre, les mardis se retrouvent généralement au milieu des deux autres groupes. Ainsi il peut être intéressant de conserver les mardis comme une classe unique.

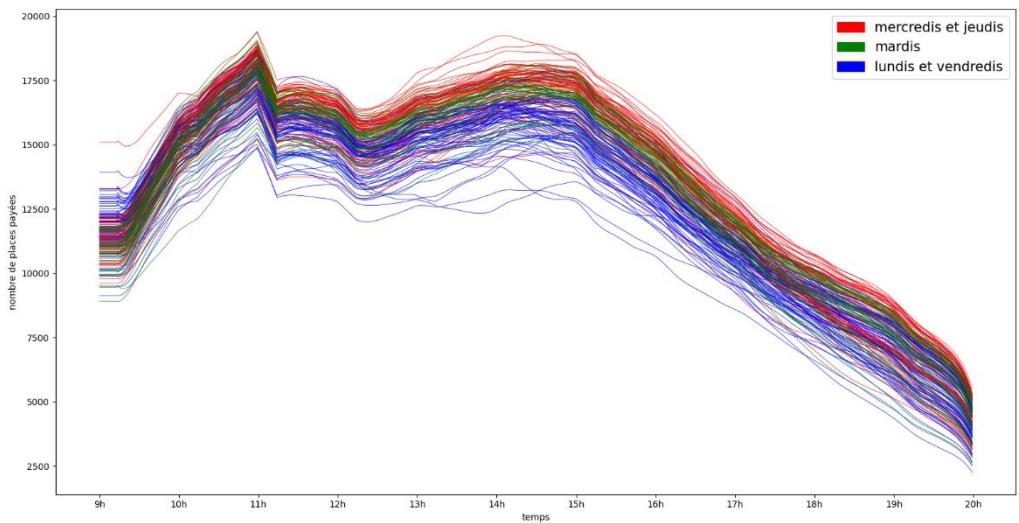


Figure 32 : Courbe des jours de la semaine divisés en trois groupes

Par ailleurs, on observe un autre découpage entre les périodes de mi-saison et les périodes de fin et début d'année qui doit être en partie liée à la nuit au vu de la pente de la courbe en fin de journée

- *Regroupement des vacances* : Les vacances scolaires ne se différencient pas des autres jours de la semaine et sont donc inclus. Hormis les vacances de janvier qui sont à part et les grandes vacances. Les mois de juillet et d'août sont regroupés en plusieurs classes, on observe les semaines d'août et les 3 samedis d'août à part. On voit que fin juillet et fin aout sont regroupés également.
- *Regroupement des samedis* : On remarque que les samedis sont regroupés ensemble sans prendre en compte la présence de petites vacances hormis les grandes vacances qui ne sont pas présente. On peut également observer des samedis de mi-saison et d'autres de début/fin d'année. La différence pourrait être liée à la nuit, comme on le voit avec les courbes.
- *Regroupement des jours de la semaine* : On observe à partir de la courbe que les lundi et vendredi sont regroupés, puis les mardis à part et enfin les mercredis et jeudis. De plus, on différencie pour chacune de ces 3 catégories les périodes de début et de fin d'année qui sont regroupées et les périodes intermédiaires regroupées elles aussi (privées des vacances d'août).

- *Exceptions* : Il existe différentes exceptions, celles qui ont été isolées par le dendrogramme (les courbes noires) et d'autres qui ont été ajoutées aux classes. On retrouve les jours de neiges, les manifestations des gilets jaunes, les ponts, et d'autres exceptions dont la raison n'a pas été expliquée comme des erreurs potentielles de données.

On analyse la variabilité au sein de chacun des jours types pour connaître la fiabilité de la classe, créer et établir une comparaison avec les jours types extraits de la méthode naïve (cf début de la partie). La variabilité de chacune des classes est calculée à partir de la distance de chacune des courbes à la moyenne de ces courbes, le calcul étant répété pour chacun des points des courbes, c'est-à-dire pour les 660 minutes de la journée (cf ci-joint). Pour chaque classe, la variabilité étant ramenée au nombre de jours en son sein, pour conserver la même échelle entre les variabilités calculées.

#### Résultats :

La variabilité moyenne dans le cas où toutes les données sont fournies brutes est de 775 places pour 15 clusters du lien moyen et de 543 places en retirant les jours exceptionnels et en conservant 15 classes. Pour la méthode naïve, la variabilité est de 1060 places pour les 5 classes établies au départ et de 868 places dans le cas où les mêmes jours exceptionnels sont retirés.

Ainsi, on observe que sans prendre en compte les zones, les 15 classes du dendrogramme permettent une plus faible variabilité et donc de meilleurs jours types que la méthode naïve.

Un dictionnaire d'occupation lissé est enregistré pour chaque zone et l'algorithme est réitéré sur celles-ci pour vérifier la bonne correspondance après division par ces zones. Le pourcentage moyen de la variabilité par zone est de 5.7 % par rapport à la taille respective de la zone et de 10 places, en utilisant la variabilité les jours types extraits à partir de l'algorithme de CAH et en retirant les jours particuliers.

Ces résultats semblent assez bons mais peuvent probablement être améliorés avec un autre algorithme. Il est intéressant de vérifier son amélioration.

$$v = \frac{\sum_{j=0}^{nbjours} \sum_{i=0}^{660} abs(courbe[i][j] - moy[i])}{nbjours}$$

	Données brutes	Jours particuliers retirés	Nombre de cluster
<b>Algorithme de CAH</b>	775	543	15
<b>Méthode naïve</b>	1060	868	5

Figure 33 : Table de comparaison des variabilités

#### g) Algorithme des k-moyennes (ou k-means)

K-means est l'algorithme de clustering le plus couramment utilisé. C'est un clustering non hiérarchique par partitionnement. La différence avec l'algorithme précédent de hiérarchie ascendante est qu'il est nécessaire de fixer le nombre de classe voulues au départ, ce qui peut compliquer la tâche. Son mode de fonctionnement varie également. [18][26][4]

##### ALGORITHME DE K-MEANS :

*Début* : Définir aléatoirement k centroïdes, k jours aléatoires sur 660 minutes. K doit être défini en amont comme le nombre de clusters voulus.

##### REPETER

1. Affecter chaque élément au centroïde le plus proche de lui en utilisant la distance de Manhattan entre chaque jour et chacun des centroïdes.
2. Recalculer le centre de chaque cluster comme un nouveau centroïde qui est le barycentre de ces jours. Soit un nouvel ensemble de 660 min correspondant à chaque instant à la moyenne des minutes pour chacun des jours associés au cluster

##### JUSQU'A

Convergence ou stabilisation (inertie). C'est-à-dire jusqu'à ce que les barycentres ne changent plus, que les clusters ne se modifient plus.

*Fin* : Récupération des différents clusters et de leur barycentre.

On récupère ainsi des clusters, mais à chaque nouveau lancement de l'algorithme, portant sur des mêmes données, les résultats sont différents. Il est donc difficile de reconnaître des jours types. Cela est dû à l'appel de départ des centroïdes aléatoires qui varient et entraîne différentes classes du fait de la proximité des données entre elles.

La moyenne de la variabilité des 20 classes créées par l'algorithme k-means sur 10 tentatives avec des valeurs de départs aléatoires différentes est de 453 places. On remarque que la variabilité est bien meilleure que celle par la CAH, mais les jours que l'on peut extraire restent similaires avec d'autres différences non identifiables.

Pour cela il est alors intéressant de mêler cet algorithme avec celui étudié précédemment.

#### h) Regroupement des deux algorithmes pour la mise en place des jours types

L'objectif en combinant les deux algorithmes est à la fois de vérifier les jours types choisis mais également de les affiner ou de les modifier en conséquence. [27]

##### - *Première option*

Tout d'abord, on retire certaines exceptions comme la neige, les manifestations ou encore certaines données très étranges. Ces jours particuliers ou exceptionnels ont d'ailleurs été trouvés grâce à l'algorithme de hiérarchie ascendante.

On relance alors l'algorithme de k-means en ayant retiré les exceptions découvertes à partir du CAH. On définit  $k$  à 15 classes et après 5 tests réalisés, on trouve une variabilité 450 places. Le nouveau résultat est donc encore plus concluant, mais toujours difficilement exploitable. En effet, les jours regroupés ne possèdent que très peu de similitude entre eux, au niveau des dates de l'années, des jours de la semaine ou encore du taux de vacances.

Une nouvelle tentative doit alors être menée par l'intermédiaire de k-means en réutilisant les jours types formés à partir de CAH.

##### - *Deuxième option*

Au lieu de générer les jours aléatoirement au début de l'algorithme, on définit les jours à partir de l'algorithme de CAH. En effet, les jours définis sont des jours choisis aléatoirement au sein de chacun des jours types provenant de CAH. Ainsi, on force les données de départs, on prend des jours aléatoirement parmi les 15 jours types du CAH et on vérifie la création des clusters, si ceux-ci suivent ceux énoncés. Si les clusters formés correspondent aux jours types de départs ainsi les jours types sont vérifiés, sinon il faut observer et comprendre les différences.

Après application dans l'algorithme, on remarque des divergences, mais pas pour tous les clusters. Certaines catégories sont reproduites presque à l'identique, comme en août ou encore les mercredis et jeudis des deux périodes, mais d'autres ne correspondent pas. Par exemple, les samedis ne suivent pas forcément de périodes (hors aout), et les mardis ne sont de nouveau pas séparés. On peut également identifier de nouveaux groupes notamment un groupe de semaine de vacances en fin juillet fin aout et janvier. Les lundis, mardis et vendredis se recoupent de manières spéciales sans trouver de raisonnement de même pour les samedis.

La nouvelle variabilité moyenne est de 438 places. On améliore ainsi la variabilité.

Il serait alors intéressant de conserver certaines de ces spécificités mais pas nécessairement les garder toutes. En effet, certaines spécificités n'étant pas analysables, on ne peut les conserver pour former des jours types de manière prédictive. Le plus important étant de pouvoir retrouver ces caractéristiques de manière prédictive et donc que les jours aient des liens retrouvables d'une année sur l'autre. Peut-être que des caractéristiques spécifiques existent mais il faudrait étudier encore plus en profondeur les données et explorer les spécificités possibles.

Les jours types conservés sont les suivants :

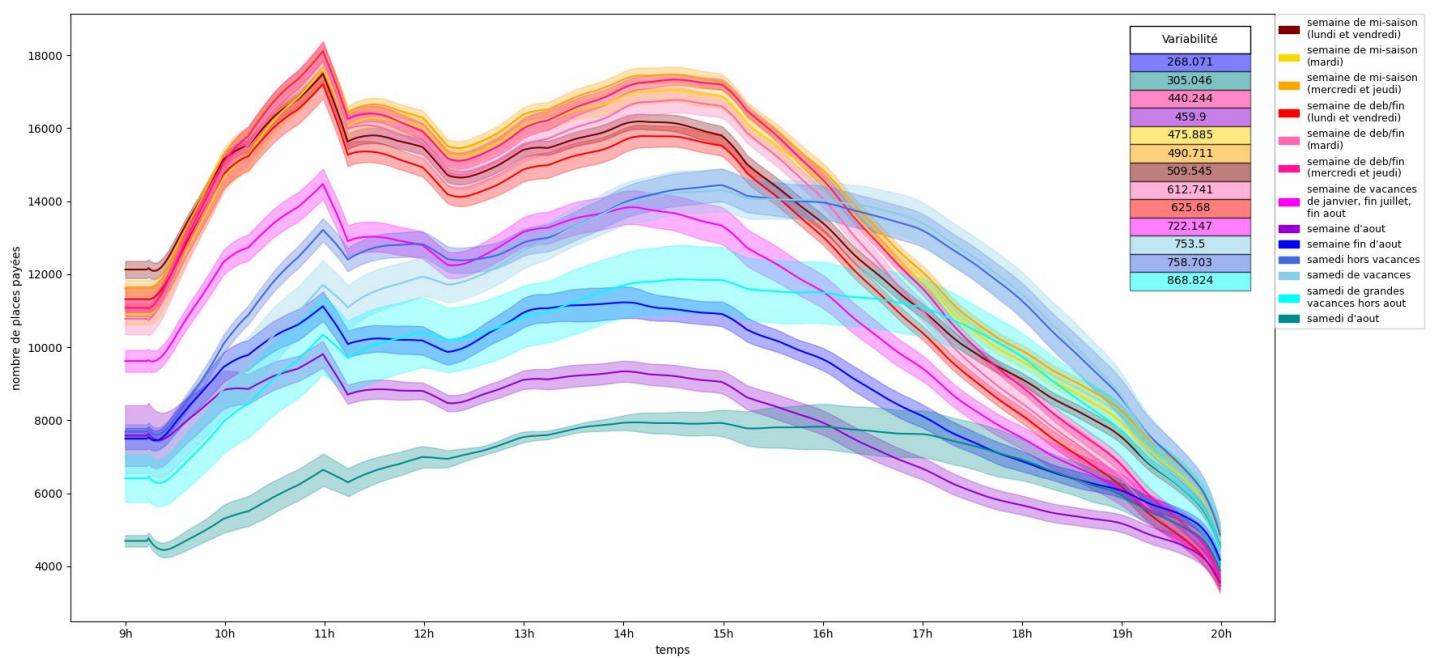


Figure 34 : Courbe des 13 jours types définitifs conservés ainsi que leur variabilité et leur intervalle de confiance à 95%

### i) Les difficultés observées

Plusieurs difficultés ont pu se présenter au cours de l'accomplissement de cette partie plutôt technique. En effet, plusieurs points se sont relevés être des challenges pour moi. Dans un premier temps, le travail de formalisation des données s'est révélé plus difficile que prévu. En effet, plusieurs difficultés se sont posées et les informations ont été découvertes au cours du processus. Par la suite, la comparaison des données deux à deux m'a semblée plus facile et très intéressante. L'utilisation des mathématiques étudiées à l'UTC m'ont servi pour comprendre certains des algorithmes et des explications mathématiques cités dans les documents qui m'ont permis l'accomplissement de ces résultats. Ensuite, la compréhension de l'algorithme de classification ascendante hiérarchique m'a paru très complexe au départ. En effet, il était nécessaire de modifier l'algorithme pour l'adapter aux données et ainsi le réaliser. De plus, pour le coder de cette manière il était nécessaire de bien prendre en main cet algorithme et le comprendre parfaitement. Cette tâche m'a pris plus de temps que ce que j'avais prévu initialement mais je suis fière d'avoir réussi à le mettre en œuvre. Finalement, la réalisation du deuxième algorithme, k-means, s'est révélée beaucoup plus facile que ce à quoi je m'attendais, car l'expérience de l'algorithme précédent m'a permis de m'adapter rapidement et de comprendre l'algorithme d'une manière très fluide, riche de mes apprentissages.

Cette partie technique, a fait entrer le projet dans le vif du sujet et l'a fait grandir fortement.

## Troisième partie : Résultats finaux et limites de cette étude

Les jours types ont été analysés et trouvés à partir des 11 mois de données reçues sur l'année 2018 à Paris. Ces jours types permettent d'évaluer une courbe d'occupation en fonction de l'heure, et ce pour chaque zone à Paris. A partir de cette courbe d'occupation, il est possible d'en déduire un taux d'occupation, en divisant par la capacité totale de la zone.

Mais ce taux d'occupation, sur lequel se basera le calcul déduisant le temps de recherche, n'est peut-être pas juste totalement car il est trouvé uniquement de par les données brutes des visiteurs sans prendre en compte d'autres paramètres comme les résidents ou encore la fraude. En regardant de plus près les données de ces visiteurs on remarque que la capacité n'est jamais atteinte et souvent très loin, c'est pour cela qu'il est nécessaire d'ajouter des caractéristiques. Mais même en cumulant ces deux critères, les données restent difficilement vérifiables, mais nous considéreront dans le cadre de cette étude, qu'elles sont juste. Ainsi, certaines limites de ces données tendent à expliquer le phénomène. Elles pourront être prise en compte dans une analyse ultérieure, avec un ajout de données ou bien de nouvelles données provenant d'une autre ville pour poursuivre le projet.

### **IV) Analyse directe du taux d'occupation et du temps de recherche**

Après avoir analysé les données en les visualisant sous forme de différents types de graphiques puis d'avoir établi des jours types, combinant plusieurs méthodologies il convient d'impliquer des facteurs annexes. En effet, pour constituer le temps de recherche d'une place sur voirie à Paris il est nécessaire de prendre en compte de nombreux paramètres, ceux-ci ayant été évoqués dans différents articles. [17]

Il est nécessaire tout d'abord de définir la manière dont les données seront présentées et retransmises, ainsi qu'identifier leur précision.

#### *Prélude : Choix de paliers d'affichage dans l'application*

Les données récoltées sont partielles car elles ne concernent que les 11 premiers mois de l'année 2018. Elles ont pour but de prédire le temps de stationnement dans le futur en s'appuyant sur des données d'horodateurs et d'applications de paiement. Ainsi, il faut garder une distance face à cette prédition et ne pas promettre aux visiteurs, futurs utilisateurs un résultat trop précis concernant leur temps de recherche.

De plus le découpage de Paris sous ces 160 zones apporte une certaine précision, mais les zones comportent tout de même environ 800 places ce qui représente un pourcentage d'erreur possible (différences possibles au sein de la même zone).

Ainsi les chiffres doivent être ronds ; afin d'offrir une certaine variabilité et un temps de recherche convenable, il apparaît judicieux de proposer des paliers de 5 minutes. D'après le CERTU, le temps de recherche maximal doit être de 20 minutes ainsi six catégories de temps de recherche sont choisies. [17] Un temps de recherche nul, dans le cas où la zone est vide, un temps de recherche négligeable, de 5 min, puis court (10 min), puis long (15 min) et très long avec 20 minutes. De plus, le dernier palier est appelé temps infini, comprenant les temps de recherche supérieurs à 20 minutes, l'objectif étant de ne plus proposer cette zone pour respecter les objectifs du CERTU.

En s'inspirant des autres études, un code couleur est instauré pour faciliter la visibilité de la disponibilité des zones, de bleu, vert, jaune, orange, rouge et noir dans le cas impossible.

### a) Taux d'occupation et temps affiliés

Le taux d'occupation est difficile à estimer et c'est la donnée la plus importante de cette étude. L'objectif des analyses précédentes consistait à l'estimer au mieux et l'objectif des suivantes correspond à l'enrichir et l'affiner.

Le taux d'occupation correspond au nombre de places occupées par rapport à la capacité totale. Ce taux est ici exprimé par zone selon chacun des jours types pour chacune des 660 minutes.

Le taux d'occupation peut être représenté par une exponentielle en fonction du temps passé, en effet plus la zone est saturée plus il sera long et fastidieux de trouver une place de stationnement.[3]

Dans les études trouvées sur le même sujet, les regroupements indiquant l'occupation sont souvent divisés en trois groupes, représentés par trois couleurs. Dans notre cas nous détaillons un dernier groupe indiquant une impossibilité de stationnement et restons plus précis avec quatre autres groupes.[8]

Dans une étude il est indiqué qu'un temps de recherche nul a lieu dans le cas où le taux d'occupation est inférieur à 60%. Il est alors considéré dans ce cas-là, que de nombreuses places se trouvent à très grande proximité du lieu d'arrivée et qu'il n'y a que le temps nécessaire au stationnement qui est comptabilisé (ce temps étant déjà pris en compte par le calculateur d'itinéraire existant).[8]

De plus, d'autres études indiquent que le temps de recherche reste négligeable si la zone est à 15% vide, soit une place sur six est alors libre. [3] Le rapport d'analyse effectué sur le temps de recherche, par Sareco, indique qu'un « taux d'occupation [...] de 85% environ, [...] correspond à un temps de recherche très court ». [17] Ce rapport indique également que le temps de recherche devient long et difficile au-dessus de 95%.

Le graphique ci-dessus indiquant le temps de recherche en fonction du taux d'occupation peut être représenté par une exponentielle. [3] D'après l'exponentielle et les données de l'article ayant découpé des zones ayant des caractéristiques très similaires aux zones de cette étude, on en extrait le tableau du temps en fonction de l'occupation à partir de l'exponentielle suivante.

$$t = 0,14 * \exp(5x)$$

avec t : temps en min

x : taux d'occupation (entre 0 et 1)

L'estimation du taux d'occupation dans cette étude est jusqu'à maintenant uniquement tirée des données brutes reçues concernant les visiteurs. Ces données étant lissée et extraites par jours types de courbes au troisième quartile, surévaluant ainsi légèrement les données pour s'assurer de leur fiabilité.

Ces données ne comptant pas les résidents, ni la fraude, la capacité maximale est largement supérieure pour presque toute les zones, ce qui ne permet pas une évaluation juste du taux d'occupation réel en 2018. C'est pourquoi il convient de réaliser des analyses supplémentaires. D'après une étude par l'Apur en 2018, le taux de vacance est estimé en moyenne à 12% à Paris avec des variations selon l'heure, le jour et la zone, ce qui pour le moment avec ces données n'est pas le cas. Nous allons donc étudier ces différents paramètres dans les parties suivantes. Mais tout d'abord, il est intéressant d'estimer le temps de marche à la suite de ce temps de recherche. [21]

Plusieurs études ont été menées sur le sujet et ce temps de marche post recherche dépend en général de ce temps-ci. Le temps de marche est donc fonction du temps utilisé en voiture lors de la recherche de place sur voirie. Or, il apparaît, selon l'étude que plus on cherche plus on s'éloigne en appliquant la méthode de la spirale, qui s'éloigne de plus en plus avec le temps. [17]

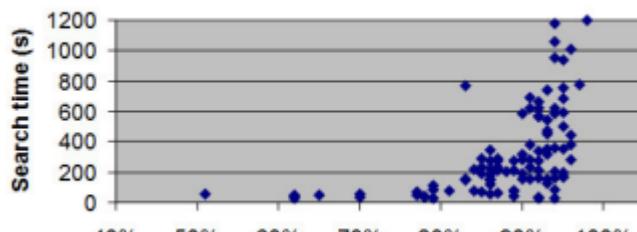


Figure 35 : temps de recherche en fonction du taux d'occupation [3][3]

Il en ressort que le temps de marche correspond à environ 60% du temps recherche. En effet, le temps de marche, est un retour en ligne approximativement droite, à pied, à partir du lieu de stationnement. En supposant, que l'on commence à chercher une place au niveau du point d'arrivée, il apparaît, d'après les différentes études menées qu'en général, les voitures vont tourner autour du point d'arrivée en s'éloignant progressivement jusqu'à former une spirale qui s'accroît de plus en plus

De plus, la vitesse moyenne de recherche sur voirie est estimée à environ 10 km/h.[3] Cela fait correspondre le temps de marche à environ 60% de ce temps de recherche, pour un retour approximativement en ligne droite (dernier rayon de recherche) à une vitesse de 4-5 km/h. Le CERTU préconise un rayon maximal de recherche de 250 mètres ce qui correspond à environ 20 minutes de recherche, la dernière grille de notre tableau ci-dessous, issu des paliers de temps de recherche formalisés au départ, des paliers d'occupation dépendant du temps par l'exponentielle et du temps de marche approximé à partir des 60% du temps de recherche.

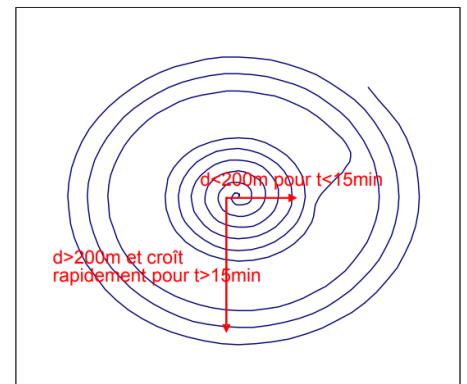


Figure 36 : La spirale du stationnement

	<b>vide</b>	<b>négligeable</b>	<b>court</b>	<b>long</b>	<b>très long</b>	<b>difficile</b>
<b>Taux d'occupation</b>	60%	75%	88%	95%	98%	100%
<b>Temps de recherche (en min)</b>	0	5	10	15	20	infini
<b>Temps de marche (en min)</b>	0	5	5	10	10	infini
<b>Code couleur associé</b>	bleu	vert	jaune	orange	rouge	noir

Figure 37 : Tableau d'occupation et de temps par catégories de couleurs

Ce tableau sera utilisé pour la suite de l'étude, afin de caractériser l'occupation. Mais avant cela il est important d'observer les autres paramètres qui pourraient modifier le taux d'occupation ou le temps liés à la recherche.

### b) Taux de renouvellement et rotation

Le taux de renouvellement représente la rotation d'une zone, combien de place sont renouvelés par rapport au nombre de places occupées. Durant une période donnée, on compte le nombre de places qui ont été libérés et reprises dans le pas et on divise tout par la moyenne de place occupées dans la zone.

Voici la formule utilisée :  $Tr = \frac{\min(\text{departs}, \text{arrivees})}{\text{occupation}}$  avec l'occupation correspondant au nombre de voiture dans la zone

Ce taux nous indique la rotation de chacune des zones pour chaque jour type, pour chacun des pas choisis et nous apporte une information supplémentaire quant à la disponibilité de la zone. Ce taux pourrait, au-dessus d'un certain pourcentage faire basculer d'une couleur vers la droite et en dessous d'un autre faire basculer d'une couleur vers la gauche, le temps de recherche.

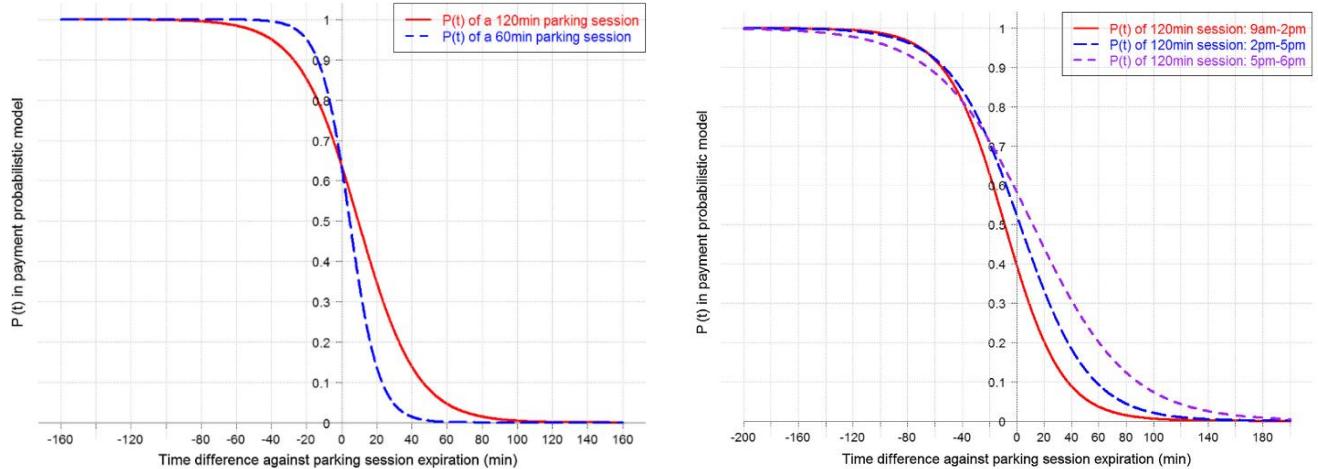
### c) Sous paiement, sur-paiement et fraude

La fraude au stationnement, c'est-à-dire le non-paiement d'un stationnement pendant une période peut avoir un important impact sur le réel taux d'occupation enregistré à partir des données de paiement. En effet, les données ne considèrent que les stationnements payés et il est souvent difficile d'estimer la réelle occupation sans capteurs sur place déterminant précisément pour chaque place son occupation dans le temps. Comme on peut le voir dans les différents rapports et études sur le sujet, c'est une donnée variable en fonction de plusieurs paramètres et difficilement estimable. Il est tout de même important de tenter une approche à ce sujet et de considérer la fraude comme une donnée centrale dans le cadre de cette étude. En effet, il a été relevé que la fraude à Paris et dans les

grandes villes est souvent plus importante, de plus les amendes à ce sujet ne sont pas toutes attribuées et de nombreux véhicules restent stationnés sans avoir payé et sans être sanctionné.

On peut estimer que les fraudeurs sont en très grande majorité des visiteurs et non des résidents au vu des prix très avantageux et des statistiques établies sur le sujet. Moins de 10 % des cas de fraudes étant des résidents à Paris. Ainsi on ajoutera la fraude uniquement sur les données liées aux visiteurs.

Dans un article réalisé sur la fraude, elle dépendrait de différents critères et varierait selon l'heure de début de fraude, le temps payé, l'heure de la journée et le jour de la semaine. D'après cet article, l'établissement du sous paiement et du sur paiement serait donc le meilleur moyen de corriger les données. [28]



**Figure 38 : Courbes issues de l'analyse de l'article pour illustrer le sur et sous paiement en fonction du temps de stationnement payé et de l'heure de la journée [28]**

Comme on le voit sur ces courbes, ces valeurs de sur paiement et de sous paiement augmentent en fonction du temps et également de l'heure de la journée, il faudrait donc se focaliser sur ces critères pour corriger les données.

Le sur paiement consiste à payer plus que le temps réel de stationnement. En effet, étant donné qu'il est nécessaire de payer à l'avance il est possible de se tromper et d'estimer un temps de stationnement inférieur au temps réel stationné. Ce cas reste plus rare et ne se compense pas avec le sous paiement, beaucoup plus présent. De plus, ce problème concerne d'avantage le paiement sur horodateur, qui représente moins de la moitié des cas de paiement à Paris, l'autre moitié provenant d'application pouvant être modifié au quart d'heure près à tout instant via son téléphone. Par ailleurs, avec les données il est impossible de l'estimer, nous poserons, dans cette étude, l'hypothèse qu'il peut être jugé négligeable. [28]

- Le sous paiement

Pour estimer le sous paiement et la fraude, certaines données de fps (amendes) ont été reçues sur la ville de Paris, au début de cette étude. Ces données se divisent en deux groupes. Tout d'abord, les personnes en sous paiement, c'est-à-dire ayant payé une partie du stationnement mais ayant été découvert au-delà de leur temps payé, dans une limite de six heures après ce temps. Mais également, des véhicules en état de fraude directe, c'est-à-dire de non-paiement de leur stationnement. Les données de sous-paiement comme de fraude sont reçus par arrondissement et s'étalent sur certains jours de l'année uniquement. Il est intéressant de les analyser pour en extraire des hypothèses.

Les données de sous paiement possèdent certaines caractéristiques, en effet, il est possible d'en extraire le temps de paiement et de récupérer un temps de fraude certain ainsi qu'un temps de fraude estimé. Il est donc intéressant de comparer les données de visiteurs ayant payés avec les données de sous paiement et de tenter d'en extraire certaines caractéristiques, notamment celles précisées dans l'article.

Tout d'abord, analysons les temps payés pour vérifier la fiabilité des données mais également observer quels temps de paiements sont plus susceptibles de se révéler être des temps de fraude. On peut ainsi vérifier si le taux de sous paiement dépend du temps payé initialement et notamment si comme dans l'article, le sous paiement

augmente avec le temps payé initialement. On corrèle les données en les ajustant les unes aux autres avec un coefficient multiplicateur selon le nombre de lignes. L'exemple a été pris sur le 12<sup>ème</sup> arrondissement.

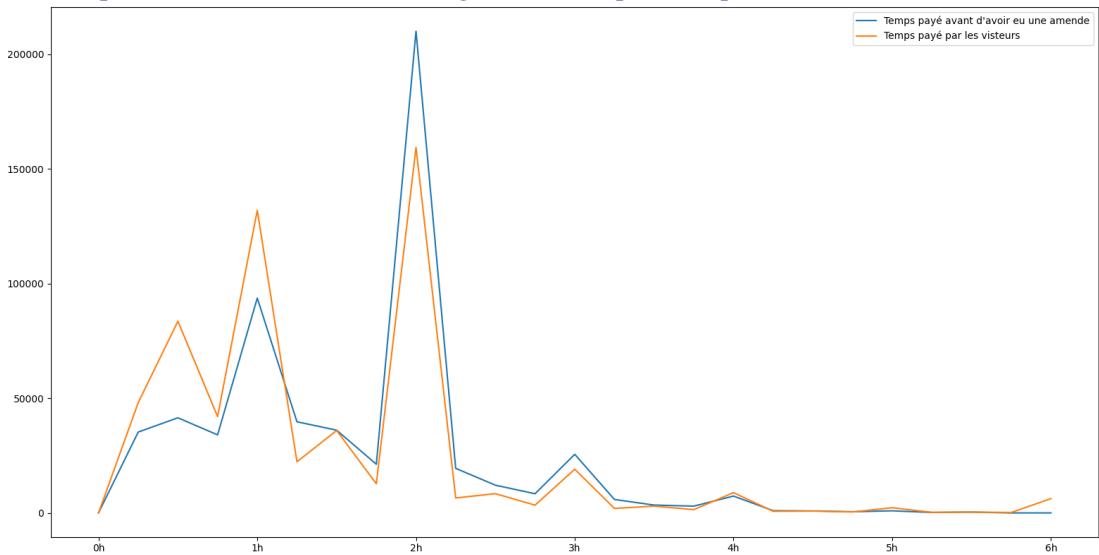


Figure 39 : Corrélation des temps de paiement avant fraude et des temps de paiements visiteurs, ajustés en fonction du nombre de données sur le 12<sup>ème</sup> arrondissement

On observe à partir de cette courbe que la proportion de temps payés par les visiteurs (tous confondus) est corrélée aux temps payés par les visiteurs contrôlés, ce qui montre une certaine fiabilité des données de fps.

En multipliant les données pour pouvoir les comparer, on observe tout de même que les conducteurs qui payent moins de temps ont tendance à moins sous payer (au-delà d'une heure de paiement), ce qui confirmerait la version considérant que les utilisateurs fraudent plus longtemps et plus souvent avec le temps. Une chose qui pourrait confirmer ce biais est le fait que l'augmentation des prix visiteur n'est pas linéaire, et que les deux premières heures sont moins chères et qu'après deux heures on a une augmentation importante du prix, ainsi cela tend à augmenter les fraudes après deux heures sur une même place, puisque les personnes auraient tendance à arrêter de payer après deux heures de paiement.

Il faut tout de même vérifier que cette correspondance n'est pas dû au fait que les sous-paiements sont simplement moins long après un faible temps de parking et donc moins probable d'être découvert et de recevoir une amende.

Analysons ensuite le temps de fraude certain pour vérifier les hypothèses :

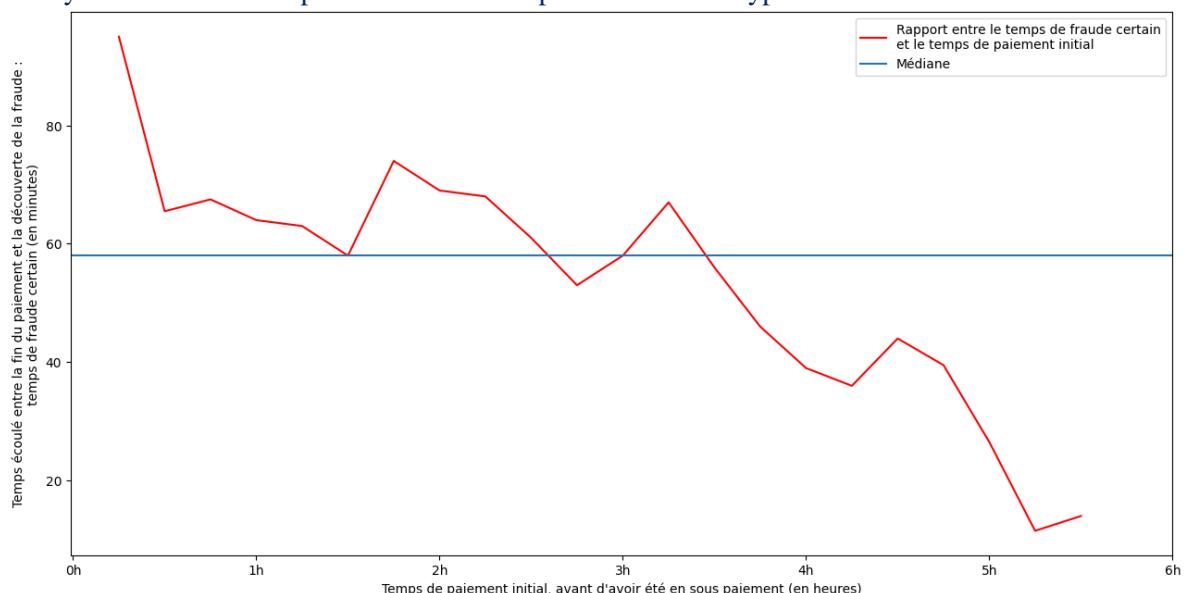


Figure 40 : Courbe du temps de fraude certain en fonction du temps initial de paiement

On observe pour chacun des temps de paiements initiaux un rapprochement autour d'une médiane. Notamment pour les paiements entre 30 minutes et 3h30, le temps de fraude étant d'en moyenne une heure. La médiane étant d'une heure également on en déduit que le temps moyen de sous-paiement à Paris est d'une heure quel que soit l'heure.

La proportion de données de sous-paiement reçues par rapport à la quantité de fraude est de 10%. On considérera le sous paiement comme 10% du pourcentage de fraude totale. On ajoutera cette proportion aux données sous la forme d'un pourcentage sur chacune des minutes, d'une heure de sous paiement.

- **La fraude**

La fraude est difficile à estimer car elle dépend de paramètres difficilement identifiables. En effet, les données de fraude reçues ne correspondent qu'aux passages des agents et peut corrompre les données en fonction d'autres paramètres non identifiés, par exemple un biais de passage dans certains lieux ou certains arrondissements plus que d'autres selon les rues. Mais également, un passage des agents selon une heure précise dans la journée ou des jours préférés. Il est ainsi impossible de formuler des jours types pour les fraudes. En effet, les jours et les horaires pouvant être corrélés aux passages des agents, aux horaires de la journée ou encore à la météo et non pas aux réelles fraudes comme l'article pouvait l'indiquer. [11]

Il serait tout de même intéressant d'estimer un pourcentage de fraude selon les arrondissements à partir des données. Pour cela, observons les données reçues et vérifions leur fiabilité. Pour chacun des 276 jours effectifs (11 mois de données, hors dimanches et jours fériés) on compte le nombre d'amende données, pour chaque arrondissement. On remarque que certains jours le passage des agents est nul ou très faible il faut donc retirer ces jours pour réaliser une bonne moyenne sur l'arrondissement. Tout d'abord, on réalise une moyenne sur les 276 jours pour chaque arrondissement et on vérifie pour chacun des jours que le nombre d'amende est bien supérieur à 10 % de cette moyenne. Ce qui permet d'éliminer les jours où les agents ne sont pas passés ou bien les jours où les données sont anormalement basses. On trouve ensuite la réelle moyenne par arrondissements en retirant ces jours, cette moyenne peut alors être analysée et multiplier par le nombre d'heure (6h) pour être comparer au nombre de paiements visiteurs en heures par jours et par arrondissements.

Néanmoins, certains jours semblent particuliers et les données ne sont peut-être pas fiables pour ceux-ci. Dans le cas où le nombre de jours retirés (anormalement faibles) est supérieur à 30% du nombre de jour total, ces jours ne peuvent être considérés comme fiable, ils ne peuvent donc pas être pris en compte pour le futur calcul de la fraude. Ci-dessous, une figure présentant les jours fiables par arrondissements. On constate que les 4 premiers arrondissements possèdent des défauts de données et ne peuvent pas être considérés comme fiables.

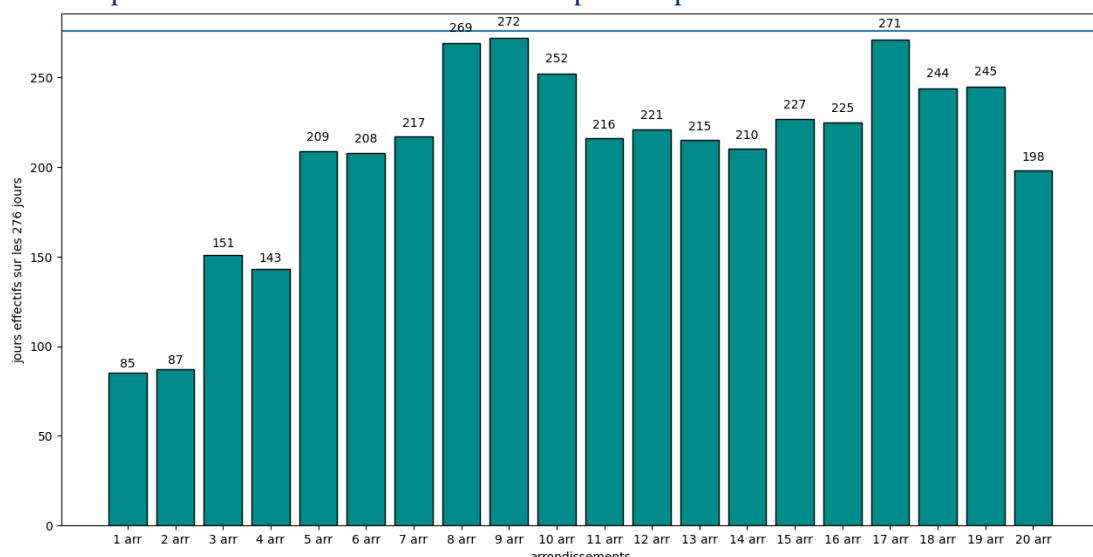


Figure 41 : Histogramme représentant le nombre de jours de données de fraude fiables reçues par arrondissements

Au vu de ces défauts de données, la solution serait donc de réaliser une régression linéaire pour établir le réel pourcentage, non plus en fonction des données reçues pour ces arrondissements mais en fonction des données

reçues pour les autres arrondissements et de trouver le nombre moyen de fraude par jours en comparant d'autres données.

Les données prises en compte pour la régression linéaire sont les suivantes : le nombre de places totales (capacité de l'arrondissement), le nombre de paiements de visiteurs par arrondissements, par minutes sur la plage horaire payante, divisée par le nombre de zones au sein de l'arrondissement et le prix de l'amende. La valeur d'entraînement étant le taux d'amendes par minutes, par arrondissements, divisée par le nombre de zones au sein de l'arrondissement. Ainsi, après avoir entraîné le modèle de prédiction avec ces données pour les 16 derniers arrondissements on relance le processus pour tous les arrondissements et on estime ainsi le nombre moyen de fraude par jours, par minutes, par arrondissements, selon le nombre de zones. A partir de ces résultats on les compare au nombre de paiements de visiteurs par jours, par minutes, par arrondissement, selon le nombre de zones et on trouve un pourcentage de fraude.

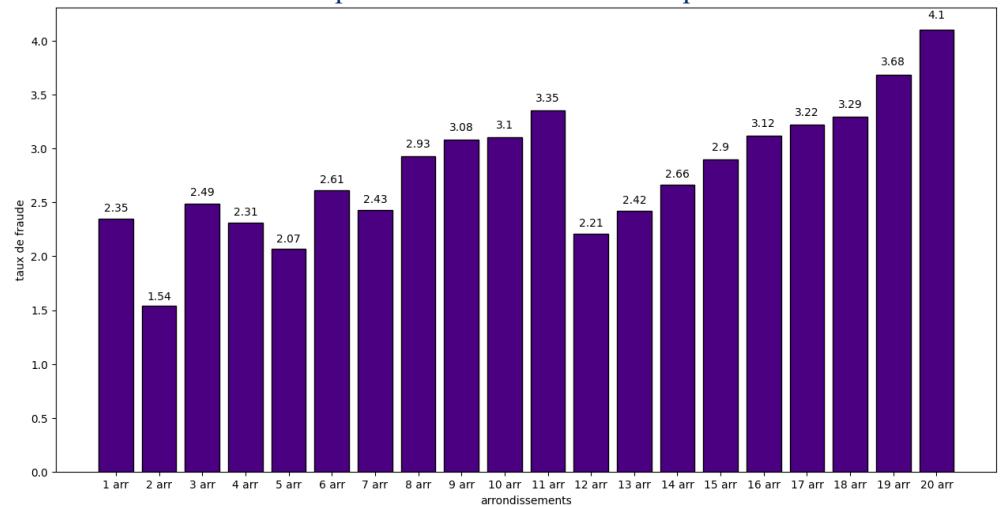


Figure 42 : Histogramme des taux de fraude par arrondissements

#### d) Densité de place par zone

La densité de place par zone est une donnée qui peut être prise en compte. En effet, de même que le taux de rotation, la densité de la zone peut jouer sur le temps de recherche. En effet, plus il y a de places à proximité les unes des autres, plus il sera rapide de trouver une place dans la zone. La taille de cette zone et le nombre de places permettent de retrouver cette densité. Ainsi, on calcule le nombre de place par zone à partir des données d'emprise et on trouve l'aire de la zone à partir des données de cartographie en json (latitude et longitude). Ces données permettent d'établir le nombre de place en fonction de la superficie en mètres carrés. Ci-dessous, on peut voir la cartographie des zones à Paris en fonction de la densité de place, le nombre indique le nombre de mètres carrés de la zone pour une place de stationnement. Ainsi plus le nombre est grand, plus il est difficile de trouver une place de stationnement.

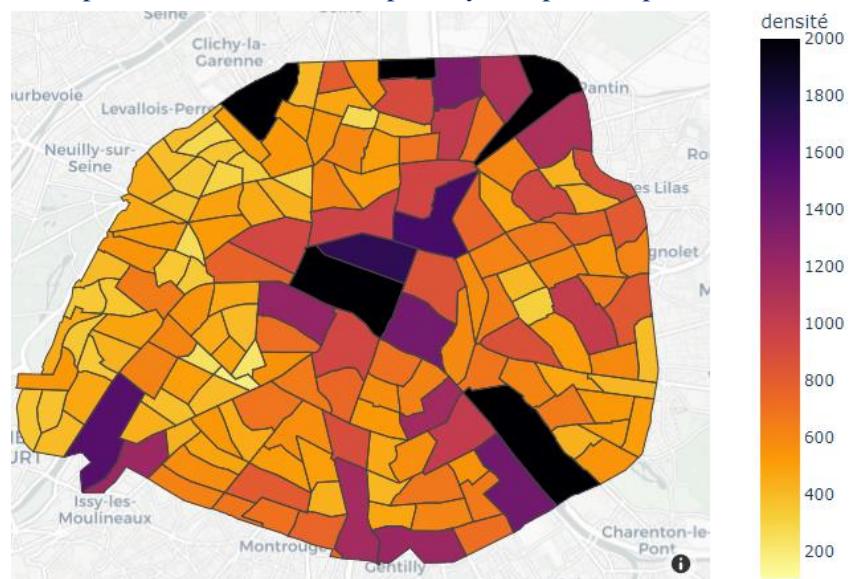


Figure 43 : Densité de places en fonction des zones

De même que pour le taux de rotation, il serait possible de prendre en compte la densité dans le cas de très faible ou très forte densité et de décaler la catégorie de couleur vers le bas ou vers le haut.

#### e) Part des résidents

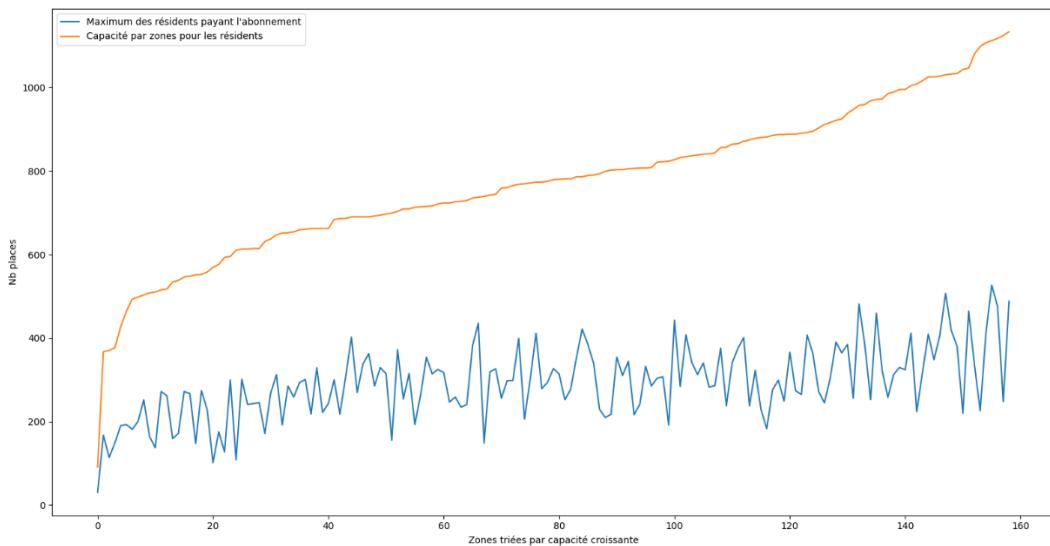
Les résidents représentent une catégorie d'usagers à part entière à Paris. Leur présence a nécessairement un impact sur les données et doit être prise en compte.

Les données concernant les résidents font partie des données reçues au départ. Elles sont exprimées par zone et s'étendent sur une période : d'une journée à une semaine. L'objectif est donc de déduire du nombre de places disponibles, l'occupation des résidents sur ces places. Mais cette occupation est difficile à obtenir au vu des données reçues. En effet, les données ne sont pas précises et ne donnent qu'une approximation du nombre de places occupées par les résidents.

Tout d'abord, les résidents ne peuvent se garer que sur un type de place, les places mixtes qui sont utilisés par les résidents et les visiteurs. Ils n'occupent que les places dans l'une des quatre zones autour de leur lieu de résidence. Lors de l'achat d'un ticket résident pour une journée (1€50) ou plus, n'indique pas le temps d'occupation réelle, il est possible que les véhicules soient ventouses ou bien pendulaires. [10] Les véhicules ventouses étant les véhicules stationnés toute la journée, tandis que les véhicules pendulaires sont les véhicules utilisés en journée et stationnés ailleurs, notamment lors de l'utilisation de la voiture pour aller jusqu'au lieu de travail. La part des résidents ventouses et des résidents pendulaires est impossible à déduire des données. Il est donc nécessaire de faire l'hypothèse que les voitures ayant payés pour une période précise resteront stationnées toute la journée à leur emplacement. En effet, il est possible que les résidents pendulaires partent avant le début de la période payante et reviennent après sa fin. On considérera que les courts déplacements des résidents en cours de journée n'impacteront que très légèrement les données, ainsi ils ne seront pas pris en compte.

Par ailleurs, il n'est pas possible de créer des jours types résidents comme cela avait été le cas pour les visiteurs, en effet, les données étant trop limitées et incertaines. Les données peuvent être prises sur une semaine entière ce qui ne permet pas de discréteriser spécifiquement des jours de l'année. De plus on imagine que les résidents sont présents toute l'année étant donné que c'est leur lieu de résidence. On ne peut également pas discréteriser le temps par minutes de la journée car les données ne précisent qu'une journée ou semaine entière.

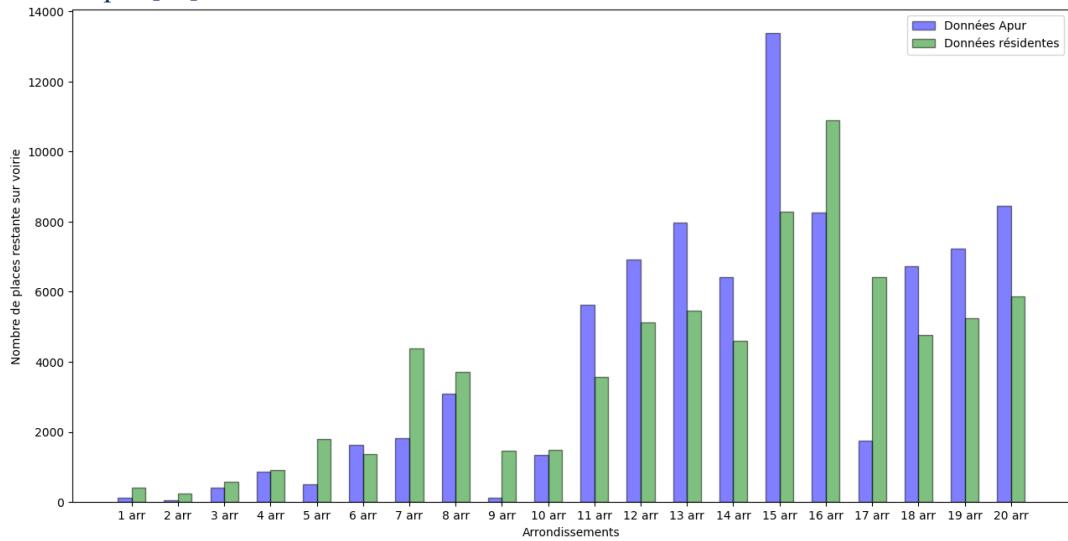
Il est donc difficile d'évaluer l'occupation réelle des résidents. Mais il est possible d'essayer de l'estimer, en considérant à partir des données le maximum d'abonnements résidents pris en même temps comme la valeur d'occupation. Cette estimation est illustrée avec le graphe suivant et fait l'hypothèse que les voitures ayant payé pour la journée sont des voitures ventouses et que ce nombre resterait fixe au cours de la journée et de l'année pour chacune des zones.



**Figure 44 : Nombre de places occupées par les résidents en fonction de la capacité des places résidentes par zones**

Il est intéressant de comparer les résultats obtenus en conservant le maximum d'abonnements résidents simultanés par zones avec les données trouvées à ce sujet dans l'étude de l'Apur (Atelier parisien d'urbanisme) ayant analysé le stationnement résident à Paris. [13] Cette analyse ne renvoie pas des résultats identiques à la réalité car elle considère tous les emplacements dans les parkings intérieurs et de logements comme occupés par une voiture résidente en décomptant le nombre de voitures à Paris par arrondissements. Elle extrait, à partir des résultats, le nombre de voitures restantes devant se garer sur la voirie et décompte ainsi les stationnements restants pour les visiteurs. Cette analyse reste imprécise également du fait que les résultats ne sont obtenus que par arrondissements.

Voici donc la différence, par arrondissements, en fonction des données et des résultats ressortant de l'analyse de l'Apur.[13]



**Figure 45 : Comparaison des données avec celles de l'Apur : places mixtes restante sur voirie après soustraction des places occupées par les résidents**

On remarque que les données de l'Apur semblent trop importantes au vu de la réalité, notamment pour les derniers arrondissements. Les résultats provenant des données, concernant l'occupation des résidents étant légèrement exagérées (notamment pour les derniers arrondissements), tandis que celles provenant de l'analyse certainement sous-évalués. Il est donc mieux de surévaluer légèrement en se basant sur des données peut être approximative que de sous-évaluer de la même manière. En effet, il vaut mieux dans le calculateur surévaluer les temps que de les sous-évaluer. [13]

#### f) Autres données qui pourraient être utilisées

D'après les recherches réalisées, plusieurs critères pourraient être intéressant à considérer pour établir au mieux, selon de nouvelles caractéristiques le taux d'occupation et le spécifier d'autant plus. [17]

- *Le nombre de places de parking en ouvrage à proximité.* Les parkings en ouvrage sont les parkings payants souterrains et dont la seule fonctionnalité est celle d'être un parking. Prendre en compte le taux d'occupation de ces parkings est une tâche beaucoup plus aisée que les stationnements sur voirie car il est uniquement nécessaire de décompter les entrées et les sorties. La prise en compte en temps réel de cette occupation, pourrait informer de l'occupation sur voirie. L'ajout de ces données dans le calculateur pourrait être intéressant.
- *Le prix du stationnement.* Le prix peut avoir un impact réel sur l'occupation, les usagers préférant en général favoriser des stationnements moins chers. Ainsi, le taux d'occupation pourrait être plus bas, dans les arrondissements où le stationnement est plus cher, les arrondissements centraux. Cette théorie semble toutefois non vérifiée, au vu des résultats obtenus, de leur plus faible capacité et de l'intérêt touristique du centre de Paris.
- *Le motif du déplacement.* La raison de la venue dans un lieu impact le stationnement, notamment d'après les études de l'Apur et les résultats parus par google maps. [21][12] En effet, le motif du déplacement, changerait le comportement de stationnement des usagers et la prise en compte du motif pourrait déduire la durée réelle de stationnement, ainsi que le temps de recherche de place. [2]
- *La file d'attente.* Il serait intéressant de tenter d'établir un modèle de file d'attente sur les données, en considérant les personnes ayant trouvé une place comme cherchant les minutes précédentes, pour en déduire le temps de recherche à chaque instant. En effet, la file d'attente consiste en un décompte du nombre de personnes cherchant une place sur un lieu donné en même temps. Etablir cette file d'attente donnerait des indications sur le nombre de personnes recherchant une place en même temps et donc la difficulté de trouver une place ou non. [15]
- *Le rayon de recherche.* Au vu du fonctionnement de recherche d'une place, sous forme d'une spirale s'étendant de plus en plus, il serait intéressant, dans les cas où le temps de recherche dépasse une certaine distance dans la spirale,

d'observer le comportement et l'occupation des zones à proximité. De plus, le comportement de ces zones pourrait influer également sur les zones à proximité, étant donné que le découpage de ces zones est virtuel. Il pourrait être alors intéressant de lisser les données sur les zones environnantes. [3]

- *La présence de bouchon.* Les bouchons de centre-ville sont souvent impactés ou créés par les personnes recherchant un stationnement. A la fois de par le ralentissement de la vitesse à environ 10 km/h pour rechercher un stationnement mais aussi à cause du temps de stationnement, une fois la place trouvée, pouvant gêner la circulation dans le cas de voies à une seule file. Ainsi la présence de bouchon pourrait donner une indication sur le nombre de personne cherchant un stationnement sur ce lieu. Il serait donc intéressant d'ajouter les bouchons en temps réel, donnée déjà présente dans le calculateur, à la recherche de stationnement.
- *La météorologie.* L'impact de la météo sur les comportements est un potentiel facteur. En effet, la météo peut modifier les habitudes, il serait alors possible d'établir d'autres jours types, en ajoutant ces données au calculateur. [6]

Après avoir défini toutes les informations qui pourraient être utilisées pour estimer le temps de recherche au sein d'une zone de Paris, il est important de se concentrer sur ce que les données actuelles permettent d'offrir. Ces informations, certaines annexes, pourraient être utilisées si de nouvelles données plus précises étaient reçues, sur la ville de Paris ou sur une autre.

## V) Présentations des résultats et limites

Les résultats obtenus tiennent compte des différents paramètres pris en compte, que ce soit les jours types acquit à partir des deux algorithmes ou encore les résidents et la fraude qui permettent d'obtenir un taux d'occupation qui peut sembler correct.

### 1) Résultats obtenus

Après ajout de la fraude sur les données des visiteurs et soustraction des résidents de la capacité on obtient un taux d'occupation pour chaque zone selon les différents jours types en divisant les courbes par la capacité définitive. Ce taux d'occupation informe sur le temps de recherche estimé. On peut observer le résultat avec par exemple la zone 16U et ses 13 jours types présentés ci-contre. On remarque les différents paliers de temps de recherche grâce aux droites les indiquant. On observe sur cette zone qu'à certains moments il arrive que le taux d'occupation dépasse légèrement les 100%, mais très peu, ce qui n'a pas un important impact sur les résultats, puisque les paliers sont approximatifs et donne une indication sur le fait la zone est saturée.

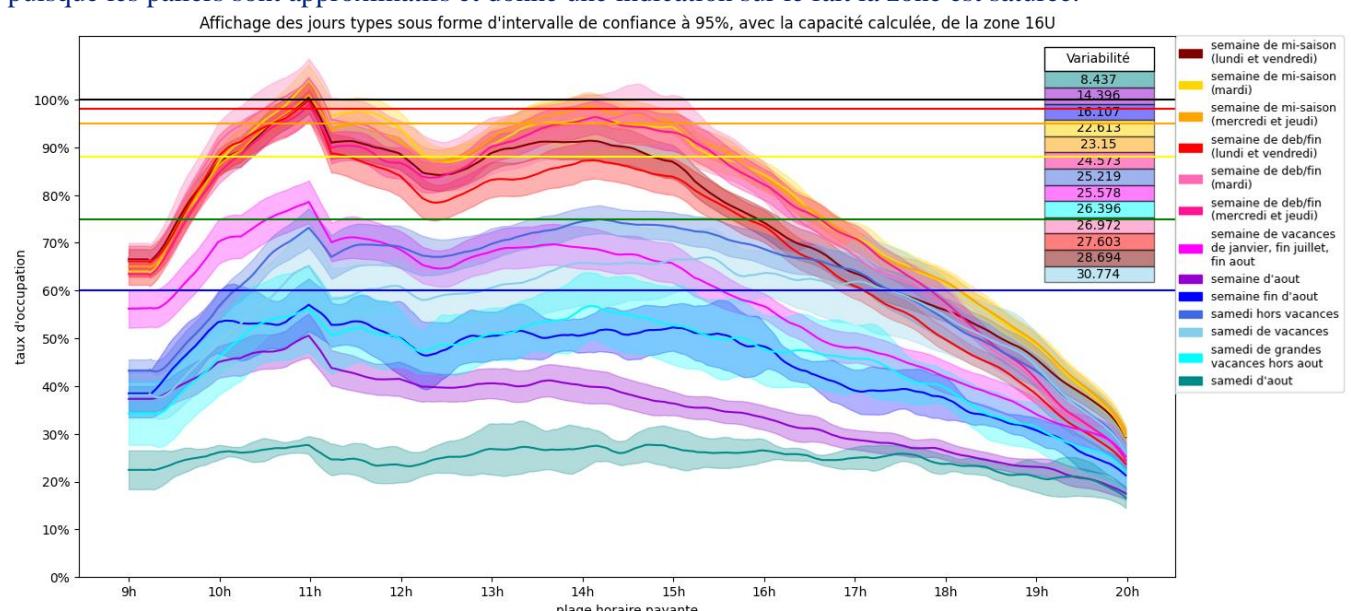


Figure 46 : Résultat d'occupation obtenu sur la zone 16U selon les jours types

Les résultats seront par la suite enregistrés, en conservant pour chacun des jours types une courbe d'occupation différente selon les zones, cette courbe correspondra au troisième quartile de toutes les courbes des jours types. Le choix de 3ème quartile permet d'éliminer les valeurs aberrantes des jours extrêmes, tout en ayant une courbe plus haute que la moyenne car il est en général mieux de surestimer le nombre de places utilisées que de le sous-estimer, notamment lorsqu'on indique un taux d'occupation à des utilisateurs.

## 2) Rendu

Les résultats ont été rendus sous deux types de forme. Un rendu des données enregistrées, utilisable dans le calculateur d'itinéraire comme occupation des places par zones et un rendu graphique pour visualiser la courbe et les données.

Le premier, correspond à un rendu sous format json des différentes particularités pour pouvoir être réutilisé directement dans le calculateur d'itinéraire ultérieurement. Plusieurs questions se sont posées quant à la forme de son rendu. Le rendu se présente sous la forme d'un fichier json comportant différentes caractéristiques et regroupant pour chaque zone, les 13 jours types et pour chacun une courbe d'occupation sous un format spécifique. En effet, pour réduire l'occupation

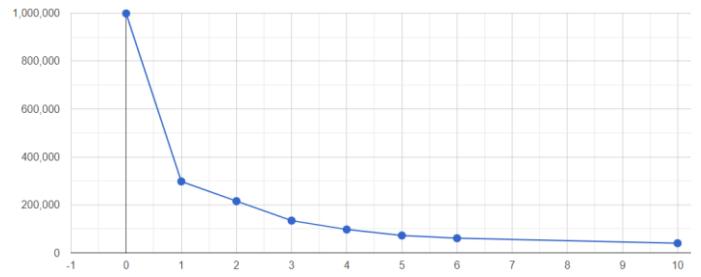


Figure 47 : Distance cumulée avec la courbe en fonction du nombre de points de rupture choisis

d'espace mémoire une solution serait de réduire la taille d'enregistrement des courbes en utilisant une régression linéaire par morceaux. En réduisant l'enregistrement des courbes on pourrait faire passer l'occupation mémoire d'une courbe de 660 états vers une courbe d'un nombre plus limité. La régression linéaire par morceaux consiste en une réduction des courbes par des droites continues.

[1] Cette approximation de la courbe s'améliore en ajoutant un nombre de points de rupture, qui correspondra

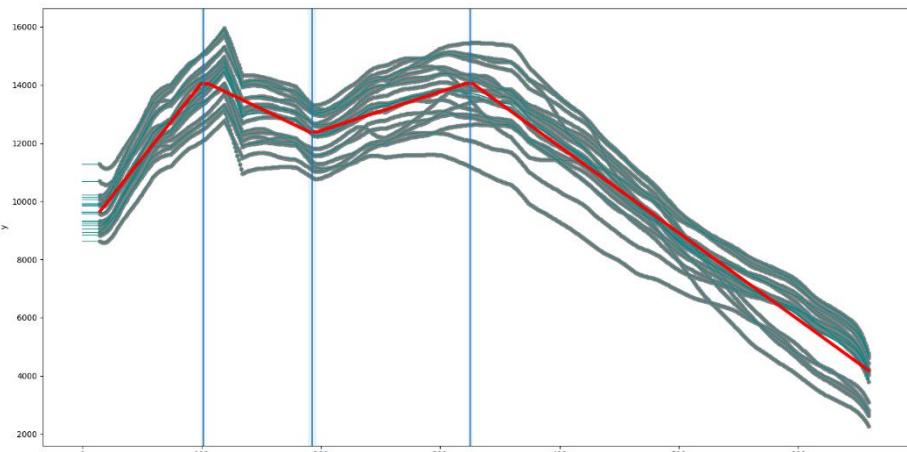


Figure 48 : Régression linéaire par morceaux sur un jour type donné avec 3 points de ruptures

au nombre de droites l'approximant. Plus le nombre de points de ruptures est grand plus la courbe est proche de la réalité. D'après la courbe ci-contre, indiquant la distance calculée entre la courbe et son approximation de régression selon le nombre de points de ruptures, on remarque que pour limiter le nombre de points enregistrés tout en conservant une approximation correcte, il serait intéressant de conserver 3 points de rupture et donc 4 droites pour approximer les courbes comme on le voit dans l'exemple ci-joint sur une jour type d'une zone donnée. Ainsi, au lieu d'enregistrer pour les 13 jours types \* les 160 zones = les 2080 courbes, 660 points d'occupation, on pourrait enregistrer les données simplifiées concernant cette courbe avec les points de départ et de fin ainsi que les coefficients directeurs des 4 lignes droites. Ce qui réduirait à 8 données. Et ferait passer l'enregistrement de 1,4 Mo à un enregistrement de 17 ko. Mais cet enregistrement présente d'autres inconvénients comme la complexification de la lecture des données dans le calculateur ce qui pourrait augmenter le temps de calcul et complexifier l'algorithme. Ainsi il semble raisonnable d'utiliser un espace mémoire d'1,4 Mo, en enregistrant pour chacun des jours types une occupation sur les 660 minutes de la journée. Une autre solution aurait été de conserver les paliers de couleur plutôt que l'occupation en indiquant la couleur d'occupation entre une période et la suivante, décomposant ainsi les données en couleurs selon des blocs de minutes successives. Cette méthode permet un gain d'espace mémoire et de lisibilité des données mais, ne permet pas de modification ultérieure selon chacun les clients

ou les usagers. En enregistrant le taux d'occupation par minutes, il est alors possible de définir les paliers de temps de recherche, de couleurs directement dans le calculateur. Le taux d'occupation est enregistré en pourcent d'occupation.

Pour retrouver les dates associées aux jours types, il suffit d'entrer un calendrier de l'année et les différentes caractéristiques correspondants aux jours types sont définis spécifiquement dans un autre fichier json, permettant alors la redirection. Les dates correspondantes aux jours types sont également enregistrées dans un fichier json pour une année, ces dates sont extraites à partir du calendrier de l'année et des jours types à partir d'un programme python. [20]

Le second rendu correspond à une carte temporelle de Paris selon une date entrée pour visualiser les données. A partir de la date entrée, le jour type correspondant à la date est retrouvé à partir du fichier json et à partir de ce jour type, il est possible de retrouver tous les taux d'occupation de chacune des zones pour toutes les minutes de la journée et d'en extraire une partie pour les rendre visible. L'affichage final est donc une carte temporelle en mouvement réalisé avec plotly (librairie python) indiquant par une couleur le temps de recherche ainsi que le taux d'occupation pour chacune des zones de Paris à une heure donnée.

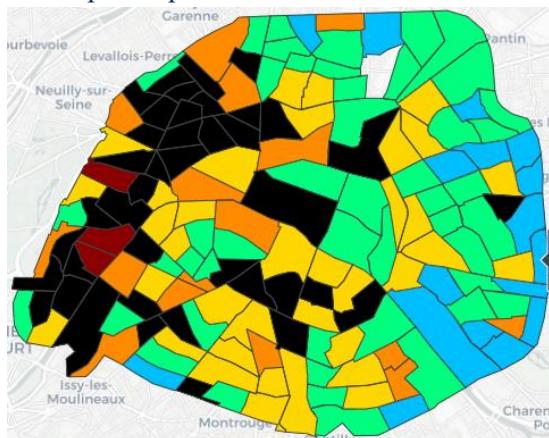


Figure 50 : carte temporelle de Paris selon le taux d'occupation des arrondissements à 14h30 sur deux jours types différents

Les deux exemples ci-dessus correspondent à l'occupation de Paris selon deux jours types différents à la même heure (14h30). Le premier correspond aux mercredis et jeudis de mi-saison (à gauche) et le second aux samedis d'août (à droite). On remarque, comme attendu, que les samedis d'août ont une occupation beaucoup plus basse par zones

### 3) Faiblesse des données

Les données reçues sont brutes, issues de données de paiements directement depuis les horodateurs et les applications. Ainsi elles peuvent présenter des défauts. Par exemple, une journée de données a dû être écartée car n'ayant aucun retour sur une journée. De plus, une zone a elle-même dû être supprimée car les données la concernant s'arrêtaient au milieu de l'année. Par ailleurs, un autre problème important est survenu

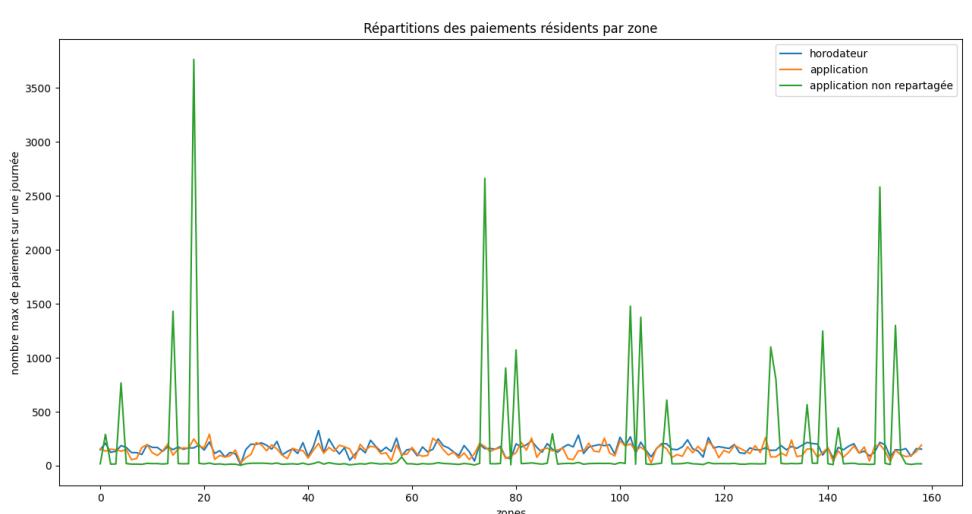


Figure 51 : Courbe illustrant les erreurs de données pour les résidents avant et après redistribution des données grâce aux zones de validité

```
{
  "Data": [
    {
      "id": "1SE",
      "arrondissement": 15,
      "DataJoursTypes": [
        {
          "IdJourType": 1,
          "Courbe": [
            94.1,
            94.1,
            94.1
          ]
        },
        {
          "IdJourType": 2,
          "Courbe": [
            94.1,
            94.1,
            94.1
          ]
        }
      ]
    }
  ]
}
```

Figure 49 : méthode d'enregistrement définitive

avec un type de paiement, PayByPhone c'est-à-dire un moyen de paiement par application, correspondant à environ la moitié des données reçues, visiteurs comme résidents. Le problème était que ce type de paiement n'ayant pas l'information concernant la zone exacte du paiement renvoyait une zone définie en fonction de l'arrondissement, par exemple pour le 1<sup>er</sup> arrondissement, la 12J était la zone utilisée. Et les réelles zones correspondant au paiement étaient situées dans la partie validityzone et indiquait pour les résidents les quatre réelles zones autour du domicile et pour les visiteurs toutes les zones de l'arrondissement. Ainsi un travail de repartage des données selon la taille de la zone a été réalisé, pour ce type de paiement. On peut voir le problème grâce à la figure ci-contre présentant les données d'applications avant et après repartage.

Une autre erreur de donnée provient des données résidentes anormalement élevées à certaines périodes de juillet et août, la raison s'est avérée être la gratuité pour les résidents les jours de pollution élevé. Or ces jours ne sont pas répertoriés et ils ont donc été très difficile à trouver.

Ces erreurs de données n'ont pas été découverte dès le début, lors de la première analyse des données, ce qui a constitué des problèmes lors de la mise en place des algorithmes ou des jours types. Ces erreurs constituent une expérience et indiquent que la première analyse des données est primordiale et doit être longue pour bien prendre en main les données et ne pas perdre de temps ensuite.

Par ailleurs d'autres erreurs de données ont été découvertes lors de la mise au propre des données et tous types d'erreurs existaient et ont dû être retirées, ce problème a complexifié fortement la tâche de formalisation des données pour les rendre analysable par la suite. Une des difficultés a également été le fait que les zones soient trop nombreuses pour être toutes analysées une par une, ce qui a complexifié la tâche de recherche des erreurs. Un des exemples peut-être la zone 34F au départ identifié comme une erreur, car ne commençant pas par un numéro d'arrondissement, mais en réalité correspondant à un mixte des arrondissements 3 et 4. Mais de nombreux autres problèmes de données ont dû être traités, au fil du temps.

#### 4) Limites

Les limites de cette étude sont variables et ont lien avec les améliorations possibles exposées dans la partie prise de recul des missions détaillées du stage.

Tout d'abord, une des limites, non abordée jusqu'ici est la différence du nombre de place entre certaines zones. En effet, par exemple certaines zones comme la 18T ne comportent que 91 places et la 11L avec 1365 places, mais ces écarts de données ne sont pas primordiaux, en effet, ils restent rares.

Une autre limite constitue l'absence de comparaison possible avec de réelles données de terrains pour vérifier la concordance des données.

Les données de fraude et de résidents pourraient être améliorés et constitue également une limite à considérer dans cette étude. Notamment, l'ajout de possible différence entre les jours de la semaine ou même des mois de l'année ainsi que des heures de la journée avec l'ajout de pendulaires ou l'augmentation du nombre de fraude en fin de journée comme évoqué dans l'article ... Les comportements des résidents pourraient notamment différer le samedi et la fraude selon le temps de stationnement voulu.

Une limite essentielle est celle de l'absence de données concernant le mois de décembre, qui pourrait avoir un impact significatif, notamment sur la formation de jours types.

Ces limites pourraient être modifiés et constituent une amélioration possible des résultats obtenus. Dans le futur, lors de l'ajout des données dans le calculateur, ces résultats ayant certaines limites présentées ci-dessus pourraient être utilisées directement dans le calculateur avec un poids plus faible ou bien modifiés au préalable avec de nouvelles données et permettant une plus solide estimation.

---

# Conclusion

Ce stage a enrichi mes connaissances l'analyse de données. J'ai dû m'approprier des données réelles en les analysant par des représentations et grâce à des statistiques. Fort de cette première analyse, j'ai pu me lancer dans la création de jours types par l'intermédiaire d'algorithmes provenant de la littérature. Les deux algorithmes utilisés étaient nouveaux pour moi. L'algorithme de classification ascendante hiérarchique puis l'algorithme de k-means m'ont permis, après une adaptation, de créer des jours type. J'ai pu associer à chaque jour type un taux d'occupation pour chaque minute de la journée. La formation de ces taux d'occupations a ensuite pu être amélioré avec l'ajout des données résidents ainsi que les données de fraude. Par cet intermédiaire, un temps de recherche de place de parking a été établi par zones. Une représentation sous forme d'une carte temporelle de couleurs a aussi été réalisé. De plus, les données d'occupations pour chaque jour type pourront être utilisées directement ou modifiées avec l'ajout de données additionnelles, dans le calculateur d'itinéraire et ainsi modifier le comportement de ce dernier. L'objectif initial concernant la prédiction de données d'occupations et de temps de stationnement par zones à Paris a donc été réalisé.

Ce stage a également été l'occasion pour moi d'approfondir mes connaissances en matière de programmation et notamment du langage python. J'ai ainsi pu mettre à profit les connaissances et les compétences acquises à l'UTC. Mon expérience à l'UTC et les UVs que j'ai pu prendre m'ont apportées dans la réalisation de mon stage. Que ce soit GE37 pour l'organisation. Ou encore MT09 et SY02 pour le lien avec les maths notamment dans l'accomplissement de certaines tâches difficile. RO03 m'a également accompagné dans ce projet pour mieux comprendre l'équipe qui m'entourait et ce qu'elle réalisait. Je me suis également servie de INF1 et de SR01 pour le lien avec le python ainsi que les regex. Mettre en concret ce que j'avais appris à l'UTC m'a permis de mieux saisir certains points qui m'avaient paru compliqué mais également de mettre en pratique ce que j'avais appris ce qui fut très enrichissant.

Ce stage se trouve être au commencement de mon choix de filière. Il constitue une bonne introduction aux métiers que je pourrais réaliser plus tard en lien avec ma future spécialisation en IAD (Intelligence Artificielle et Données). L'apprentissage de ce nouveau métier m'a beaucoup plu.

Ce stage m'a aussi permis de me confronter aux réalités de l'entreprise et l'apprentissage du travail au sein d'une équipe. J'ai aussi pu observer le travail de l'ingénieur dans l'entreprise et les tâches de programmation réalisable.

J'ai beaucoup apprécié ce stage de 6 mois au sein de Cityway et de son équipe TripPlanner. J'ai ressenti une ambiance conviviale entre les équipes de cette entreprise. J'ai appris de nombreuses choses à la fois sur le plan technique mais également sur le plan humain. J'en retiens de nombreuses choses pour mes futurs projets et mes nouvelles expériences.

---

# Annexes

## Table des matières détaillée

Remerciements .....	1
Résumé technique.....	2
Introduction.....	3
Présentation de Cityway.....	4
1) Un peu d'histoire	
2) Sa mission	
3) Ses produits	
4) L'évolution de sa stratégie	
5) Son processus	
6) Son organisation	
7) L'équipe TripPlanner	
Les missions du stage .....	12
1) Le sujet	
2) Le planning	
3) Les contributions	
4) Les outils et technologies	
5) Prise de recul	
Etude du stationnement sur voirie.....	19
Première partie : Recherches, prises d'informations et analyses.....	19
I) Recherches pratiques	
1) Stationnement sur voirie	
2) Amendes	
a) Le fonctionnement	
b) Les chiffres	
3) Résidents	
4) Recherches d'articles scientifiques	
II) Première analyse des données et statistique	
a) Les données	
b) Premières statistiques et analyses	
Deuxième partie : Analyse technique des données et algorithmes.....	25
III) Création des jours types	
a) Méthode naïve	
b) Formalisation des données d'entrées	

---

c) Courbe du nombre de paiements visiteurs sur la journée	
d) Choix de la distance de comparaison	
e) Algorithme de classification hiérarchique ascendante	
f) Dendrogrammes et courbes associées	
g) Algorithme des k-moyennes (ou k-means)	
i) Regroupement des deux algorithmes pour la mise en place des jours types	
j) Les difficultés observées	
Troisième partie : Résultats finaux et limites de cette étude.....	38
<b>IV) Analyse directe du taux d'occupation et du temps de recherche</b>	
Prélude	
a) Taux d'occupation et temps affiliés	
b) Taux de renouvellement et rotation	
c) Sous paiement, sur-paiement et fraude	
d) Densité de place par zone	
e) Part des résidents	
f) Autres données qui pourraient être utilisées	
<b>V) Présentations des résultats et limites</b>	
1) Résultats obtenus	
2) Rendu	
3) Faiblesse des données	
4) Limites	
<b>Conclusion .....</b>	51
<b>Annexes .....</b>	52

# Table des illustrations

Figure 1 : Produit MaaS de Cityway.....	5
Figure 2 : Format du processus de mise en production d'un projet.....	6
Figure 3 : Zoom sur la partie techniques et les différentes équipes qui s'y trouvent.....	7
Figure 4 : Sélection des contraintes préférentielles par l'usager sur le site Fluo .....	8
Figure 5 : Extrait du site Fluo, calcul d'itinéraire entre Paris et Nancy à 10h15 (solution transport en commun)9	9
Figure 6 : Extrait du site Fluo, calcul d'itinéraire entre Paris et Nancy à 10h15 (solution voiture et vélo).....	9
Figure 7 : Fonctionnement technique du calculateur d'itinéraire à Cityway .....	10
Figure 8 : Temps de réponse moyen selon les modes cherchés par le calculateur d'itinéraire, outil : Grafana	11
Figure 9 : Planning initial des tâches à réaliser .....	14
Figure 10 : Planning GANTT final du stage avec les dates de réalisation des principales tâches et les remises de rapports .....	14
Figure 11 : OBS présentant ma place au sein de Cityway durant ce stage .....	17
Figure 12 : Les 160 zones de stationnement à Paris.....	19
Figure 13 : Tarif du stationnement visiteur .....	20
Figure 14 : Augmentation non linéaire du tarif visiteur en fonction des zones .....	20
Figure 15 : Camembert des différents régimes de places existants à Paris .....	22
Figure 16 : Capacité des zones de stationnement.....	22
Figure 17 : Proportion de chaque profil utilisateur dans les données reçues .....	23
Figure 18 : Maximum d'occupation des visiteurs en simultané sur les 11 mois de l'année pour chacune des zones, en comparaison avec la capacité de ces zones .....	23
Figure 19 : Courbes d'occupation simultanée, selon les types de profils, sur deux semaines .....	24
Figure 20 : Nombre de stationnements payés, par les visiteurs, par paliers de 15 minutes.....	24
Figure 21 : Visualisation chronologique des plages payantes et gratuites .....	27
Figure 22 : Explication de deux manières de lire une date .....	27
Figure 23 : Figure illustrant le décompte des places occupées en simultané, dans le temps.....	28
Figure 24 : Courbe des visiteurs simultanés par jours lissés sur 15 minutes.....	29
Figure 25 : Courbe d'occupation des visiteurs par jours.....	29
Figure 26 : Comparaison des jours deux à deux par distance de Manhattan, les 333 jours sont affichés dans l'ordre chronologique.....	30
Figure 27 : Représentation de l'algorithme sous forme d'un dendrogramme .....	31
Figure 28 : Format des données de clustering .....	31
Figure 29 : Explication de la méthode de regroupement par distance des barycentres.....	32
Figure 30 : Dendrogramme issu de l'algorithme d'ascendance hiérarchique avec lien moyen.....	33
Figure 31 : Courbes avec couleurs associées aux dendrogrammes ci-dessus .....	33
Figure 32 : Courbe des jours de la semaines divisés en trois groupes .....	34
Figure 33 : Table de comparaison des variabilités .....	35
Figure 34 : Courbe des 13 jours types définitifs conservés ainsi que leur variabilité et leur intervalle de confiance à 95%.....	37
Figure 35 : temps de recherche en fonction du taux d'occupation [3] .....	39
Figure 36 : La spirale du stationnement.....	40
Figure 37 : Tableau d'occupation et de temps par catégories de couleurs .....	40
Figure 38 : Courbes issues de l'analyse de l'article pour illustrer le sur et sous paiement en fonction du temps de stationnement payé et de l'heure de la journée [28].....	41

---

Figure 39 : Corrélation des temps de paiement avant fraude et des temps de paiements visiteurs, ajustés en fonction du nombre de données sur le 12ème arrondissement.....	42
Figure 40 : Courbe du temps de fraude certain en fonction du temps initial de paiement .....	42
Figure 41 : Histogramme représentant le nombre de jours de données de fraude fiables reçues par arrondissements .....	43
Figure 42 : Histogramme des taux de fraude par arrondissements .....	44
Figure 43 : Densité de places en fonction des zones.....	44
Figure 44 : Nombre de places occupées par les résidents en fonction de la capacité des places résidentes par zones .....	45
Figure 45 : Comparaison des données avec celles de l'Apur : places mixtes restante sur voirie après soustraction des places occupées par les résidents.....	46
Figure 46 : Résultat d'occupation obtenu sur la zone 16U selon les jours types.....	47
Figure 47 : Distance cumulée avec la courbe en fonction du nombre de points de rupture choisis.....	48
Figure 48 : Régression linéaire par morceaux sur un jour type donné avec 3 points de ruptures.....	48
Figure 49 : méthode d'enregistrement définitive .....	49
Figure 50 : carte temporelle de Paris selon le taux d'occupation des arrondissements à 14h30 sur deux jours types différents .....	49
Figure 51 : Courbe illustrant les erreurs de données pour les résidents avant et après redistribution des données grâce aux zones de validité .....	49

---

## Glossaire

ADEME	Agence de l'Environnement et de la Maîtrise de l'Energie
API	Application Programming Interface / interface de programmation d'application
Apur	Atelier parisien d'urbanisme
A*	Algorithme de recherche de chemin dans un graphe
CERTU	Centre d'études sur les réseaux, les transports, l'urbanisme
CSA	Connection Scan Algorithm
Dataframe	Tableaux à plusieurs dimensions permettant le traitement d'analyse de données en informatique (présent dans une librairie python notamment)
Djikstra	Algorithme de résolution du problème de plus court chemin
EGT	Enquête Globale des Transports réalisée en 2010 par des sondages sur les Franciliens
Emprise	Bloc de stationnements sur voirie
FPS	Forfait Post-Stationnement, amende en cas de non-paiement du stationnement
Intermodal	Combine différents modes de transport
Jours types	Jours de référence ayant des caractéristiques établies pouvant être précisées et retrouvées
Json	JavaScript Object Notation : format de données textuelles
MaaS	Mobility as a Service
Multimodal	Retour de plusieurs résultats avec des moyens de transports différents
OBS	Organization Breakdown Structure : organigramme de gestion de projet
Pendulaires	Type d'usager utilisant la voiture pour aller au travail la journée et revenant se garer sur la voirie proche de leur résidence en soirée
Projet Mi2	Projet avec la ville de Paris d'où proviennent les données
QA	Quality Assurance/ Assurance qualité
RI	Recherche d'Itinéraire
SISMO	Système Intégré de Services à la mobilité de l'Oise.
Sous paiement	Paiement inférieur au temps d'occupation réel
Sur paiement	Paiement supérieur au temps d'occupation réel
Voirie	Rue

## Références

[1]	Alfonso Croeze, L. P. (2012). Nonlinear Least-Squares Problems with the. LSU&UoM.
[2]	Bayardin, V., & Jabot, D. (2012, 23 septembre). <i>En Île-de-France, la moitié des actifs parcourent plus de neuf kilomètres pour aller travailler - Insee Flash Ile-de-France - 60</i> . Accueil – Insee – Institut national de la statistique et des études économiques   Insee. <a href="https://www.insee.fr/fr/statistiques/5425974#tableau-figure3">https://www.insee.fr/fr/statistiques/5425974#tableau-figure3</a>
[3]	Belloche, S. (2015). On-street parking search time modelling and validation with survey-based data. <i>Transportation Research Procedia</i> , 6, 313-324.
[4]	Benzaki, Y. (2018, 10 avril). <i>Tout ce que vous voulez savoir sur l'algorithme K-Means</i> . Mr. Mint : Apprendre le Machine Learning de A à Z. <a href="https://mrmint.fr/algorithme-k-means">https://mrmint.fr/algorithme-k-means</a>
[5]	Black, P. E. (2006). Manhattan distance"" Dictionary of algorithms and data structures. <a href="http://xlinux.nist.gov/dads/">http://xlinux.nist.gov/dads/</a> .
[6]	<i>Climatologie de l'année 2018 à Paris-Montsouris - Infoclimat</i> . (2018). Infoclimat - la météo en temps réel : observations météo en direct, prévisions, archives climatologiques, photos et vidéos... <a href="https://www.infoclimat.fr/climatologie/annee/2018/paris-montsouris/valeurs/07156.html">https://www.infoclimat.fr/climatologie/annee/2018/paris-montsouris/valeurs/07156.html</a>
[7]	<i>Enquête Globale des Transports</i> . (s. d.). OMNIL. <a href="https://omnil.fr/spip.php?article227">https://omnil.fr/spip.php?article227</a>
[8]	<i>Etude de stationnement Chinon centre-ville</i> . (2019). Chinon: Sareco.
[10]	Gantelet, E. (2012). <i>Concevoir un plan de stationnement</i> . Sareco.
[11]	GART et CEREMA. (2019). <i>Réforme du stationnement payant sur voirie : bilan de la première année de mise en oeuvre</i> .
[12]	Google Maps. (2022). <i>Rapports sur la mobilité de la communauté - COVID-19</i> . Google Maps.
[13]	HANAPPE, F., LO PINTO, A., & VAULÉON, Y.-F. (2019). <i>Evolution du stationnement et nouveaux usages de l'espace public Volet 1</i> . Paris: Apur.
[14]	<i>Home — Paris Data</i> . (s. d.). Home — Paris Data. <a href="https://opendata.paris.fr/pages/home/">https://opendata.paris.fr/pages/home/</a>
[15]	Jordon, D., Hampshire, R., & Fabusuyi, T. (2021). Estimating parking occupancy using smart meter transaction data. <i>arXiv preprint arXiv:2106.02270</i> .
[16]	<i>La classification ascendante hiérarchique – lemakistatheux</i> . (2016, 23 juin). lemakistatheux. <a href="https://lemakistatheux.wordpress.com/category/outils-danalyse-non-supervisee/la-classification-ascendante-hierarchique/">https://lemakistatheux.wordpress.com/category/outils-danalyse-non-supervisee/la-classification-ascendante-hierarchique/</a>
[17]	Lefauconnier, A., & Gantelet, E. (2005). Le temps de recherche d'une place de stationnement. <i>Prédit. ADEME SARECO</i> .
[18]	MacQueen, J. (1967, June). Classification and analysis of multivariate observations. In <i>5th Berkeley Symp. Math. Statist. Probability</i> (pp. 281-297). Los Angeles LA USA: University of California.
[19]	Mannini, L., Cipriani, E., Crisalli, U., Gemma, A., & Vaccaro, G. (2017). On-street parking search time estimation using fcd data. <i>Transportation Research Procedia</i> , 27, 929-936.
[20]	Ministère de l'Éducation Nationale et de la Jeunesse. (s. d.). <i>Le calendrier scolaire - http://data.gouv.fr</i> . Accueil - <a href="http://data.gouv.fr">http://data.gouv.fr</a> . <a href="https://www.data.gouv.fr/fr/datasets/le-calendrier-scolaire/#resources">https://www.data.gouv.fr/fr/datasets/le-calendrier-scolaire/#resources</a>
[21]	NICOL, M.-A., & LO PINTO, A. (2019). <i>Analyse des mobilités domicile-travail Volet 1</i> . Apur.
[22]	Nielsen, Frank. (2016). Hierarchical Clustering. 10.1007/978-3-319-21903-5_8.
[23]	Sonntag, J., Engel, M., & Schmidt-Thieme, L. (2021, May). Predicting Parking Availability from Mobile Payment Transactions with Positive Unlabeled Learning. In <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> (Vol. 35, No. 17, pp. 15408-15415).

[24]	<i>Stationnement.</i> (s. d.). <a href="http://Paris.fr">http://Paris.fr</a> , site officiel de la Ville de Paris. <a href="https://www.paris.fr/stationnement">https://www.paris.fr/stationnement</a>
[25]	Tamrazian, A., Qian, Z., & Rajagopal, R. (2015). Where is my parking spot? Online and offline prediction of time-varying parking occupancy. <i>Transportation Research Record</i> , 2489(1), 77-85.
[26]	Vialle, S. (2019). Big Data : Informatique pour les données et calculs massifs - Algorithmes de clustering. CentraleSupélec.
[27]	Vieille, M.-J. (s. d.). <i>Comment faire quand la CAH est dépassée ? - Lovely Analytics</i> . Lovely Analytics. <a href="https://www.lovelyanalytics.com/2017/11/18/cah-methode-mixte/">https://www.lovelyanalytics.com/2017/11/18/cah-methode-mixte/</a>
[28]	Yang, S., & Qian, Z. S. (2017). Turning meter transactions data into occupancy and payment behavioral information for on-street parking. <i>Transportation Research Part C: Emerging Technologies</i> , 78, 165-182.
[29]	Zhang, J. (2007). <i>Visualization for information retrieval</i> (Vol. 23). Springer Science & Business Media.
[30]	<i>10 500 FPS par jour à Paris en 2018.</i> (2019, 3 février). Amende Forfait Post Stationnement. <a href="https://fps-stationnement.fr/actualite/10-500-fps-par-jour-a-paris-en-2018-1518/">https://fps-stationnement.fr/actualite/10-500-fps-par-jour-a-paris-en-2018-1518/</a>