

London Fever



Analyse des temps de réponse
des pompiers londoniens

BootCamp Data Analyst

Smail MAKOUDI

Charlie PARÉ

Roman VUILLAUME

London Fever

Analyse des temps de réponse
des pompiers londoniens

by

Smail MAKOUDI
Charlie PARÉ
Roman VUILLAUME

Mentor: Greg Tordjman
Project Duration: April, 2021 - May, 2021

Cover Image: Entraînement des pompiers londonien, image provenant de la communication de la London Fire Brigade



DataScientest.com

List of Figures

1.1	London Fire Brigade' Stations	5
2.1	étude de la répartition des temps d'intervention	7
2.2	évolution du temps d'intervention moyen et son écart type	8
2.3	Part des interventions par heures de la journée	8
2.4	Temps d'intervention moyen en fonction de la distance relative avec le centre de Londres après 2016	9
2.5	Interventions par station	10
2.6	Carte des distances moyennes par station	11
2.7	évolution du temps d'intervention moyen et son écart type	11
3.1	Distribution des temps de réponse en fonction de la distance	12
3.2	Représentation géographique des incidents coloré par station ayant participé à l'intervention.	13
3.3	Résultats de la classification par les K moyens, en fonction du fragment du segment utilisé pendant l'entraînement.	14
3.4	Comparaison des performances GMM, K moyens	15
3.5	Projection géographique des GMM et de leurs covariances	16
4.1	Illustration des différentes méthodes de calcul d'itinéraire, avec le trajet calculé, la distance totale du trajet, et le temps de calcul. La distance euclidienne est plus importante que les trajets car elle est calculée sur la projection UTM des coordonnées, tandis que les trajets calculés avec OSMNX et PBF calcul sur les coordonnées avant de réaliser la projection, et accumulent donc moins d'erreur.	18
4.2	Relation entre distances euclidiennes et distances calculées avec PBF	19
4.3	Résultats final.	20

List of Tables

1.1	Variables sélectionnées pour l'étude des temps de réponses des pompiers londoniens.	3
A.1	Station's name and coordinates used in this study.	23

Introduction

*"Everybody's got the fever
That is something you all know
Fever isn't such a new thing
Fever started long ago"*

Elvis Presley, Fever, 1960

Créée il y a presque deux siècles, la *London Fire Brigade* (LFB) intervient sur les situations d'urgences dans toute l'aire urbaine de Londres. Elle est composée de 102 casernes de pompiers et d'une brigade fluviale.

Le temps de réponse des stations au signalement d'une urgence est un facteur majeur pour la limitation des dégâts à la fois physique et matériel. La LFB s'est donc fixé comme objectif l'arrivée des premiers secours dans un temps inférieur à 6 minutes (360 sec) en moyenne, et d'une seconde équipe en assistance, si nécessaire, dans les 8 premières minutes (480 sec) après le signalement¹.

Nous allons nous intéresser à ce temps de réponse, plus particulièrement au temps nécessaire pour les services de secours pour intervenir une fois mobilisé. Pour cela nous étudierons la vaste quantité de données mises à disposition par la LFB². Il sera notamment évoqué comment la donnée y est structurée, et de la mettre en perspective pour l'analyser.

La LFB met non seulement à disposition ses propres données, mais également des analyses de ces dernières³. C'est pourquoi nous nous essayerons d'enrichir leurs approches avec des analyses et des données complémentaires. Pareillement, nous essayerons de prédire quelles stations seraient mobilisées sur un incident à l'aide d'un algorithme entraîné sur une partie des données. A l'aide de cette prédiction, nous pourrons estimer le temps d'une intervention en se basant sur l'historique de la station.

¹<https://data.london.gov.uk/dataset/lfb-financial-and-performance-reporting-2019-20>

²<https://data.london.gov.uk/publisher/lfb>

³voir notamment le dasboard Power BI mis à disposition à cette adresse <https://data.london.gov.uk/dataset/london-fire-brigade-incident-records>

Étude du Dataset

1.1. Présentation du jeu de données

Les données mises à disposition par la LFB sont massives. Concaténées, elles représentent un peu moins d'un gigaoctet au format csv. Elles sont mises à disposition sous la forme de deux set séparant les mobilisations et les incidents. Le *dataset* Mobilisation contient 2.032.480 entrées. Celui concernant les incidents contient 1.864.458 entrées. Celle-ci couvrent toutes les interventions des unités de secours depuis le début de l'année 2009 jusqu'à la fin du mois de janvier 2021. Les deux sets de données cumulent 75 colonnes individuelles.

Ces colonnes contiennent, entre autres, les données suivantes¹ :

En plus de cet échantillon de colonnes, explicité par le service data de la LFB, le jeu de données contient un large éventail de détails complémentaires dont les plus important pour notre étude sont :

- les colonnes *Easting* et *Northing* qui contiennent les coordonnées des interventions selon un système de coordonnées géographiques cartésiennes.
- les colonnes *DateOfCall*, *CalYear*, *HourOfCall* et *TimeOfCall* qui contiennent la date, l'année, l'heure et la date complète (jour, mois, année et heure) de l'intervention

Le jeu de donnée est brut, contenant parfois de très grands nombres de valeurs *null* dans certaines colonnes, comme par exemple celle *DelayCodeID* et *DelayCode_Description*. Mais, comme nous le verrons dans la partie suivante, la plupart des valeurs *null* ont en partie été écartées par les manipulations de la base de données.

On notera, détail cocasse, que certaines variables comme *DataOfCall* ont très probablement été entrées par un francophone entre 2009 et 2013. En effet, certains mois étaient non pas enregistrés sous un format numérique (ex: 01-06-2011) mais en pleines lettres et en français (ex: 01-juin-2011).

¹Merci à Jenny Taylor, responsable de la donnée de la LFB pour les informations détaillées

Table 1.1: Variables sélectionnées pour l'étude des temps de réponses des pompiers londoniens.

Id	Colonne	Type des données
0	IncidentNumber	Identifiant unique de la LFB pour chaque incident
1	ResourceMobilisationID	Identifiant unique de la LFB pour la mobilisation
2	Resource_Code	Code des outils mobilisés
3	PerformanceReporting	Précise si cet incident a été utilisé pour les rapports de performance
4	TimeMobilised	L'heure et la date de mobilisation
5	TimeMobile	L'heure et la date de mobilité des forces de secours
6	TimeArrived	L'heure et la date d'arrivée sur le lieu de l'incident
7	AttendanceTimeSeconds	Différence en seconde entre le temps de départ et d'arrivée
8	TimeLeft	L'heure à laquelle les secours ont quitté le lieu d'intervention
9	TimeReturned	L'heure de retour des secours à leur station
10	DeployedFromStation_Code	Code de la station dont les forces ont été déployées
11	DeployedFromStation_Name	Nom de la station dont les forces ont été déployées
12	DeployedFromLocation	Indique si le matériel était déployé de sa base originelle ou non
13	MobilisationOrder	Ordre dans lequel le matériel a été mobilisé
14	PlusCode_code	Code indiquant si la mobilisation était partie de la PDA(Pre-Determined Attendance) initiale, les secours, etc.
15	PlusCode_Description	Descriptif du code précédent
16	DelayCodeID	ID correspondant à toutes circonstances ayant entraîné un potentiel retard dans l'intervention
17	DelayCode_Description	Descriptif de l'ID précédente

1.2. Nettoyage et enrichissement du jeu de données

1.2.1. Première approche du jeu de données

En raison de l'importance du jeu de données il a été décidé très tôt dans le processus d'analyse d'écarter plusieurs colonnes dont la pertinence semblait faible au regard de la problématique. Par exemple des colonnes précisant le coût de l'intervention, ou le code du matériel mobilisé. Pour ces raisons, 60 variables ont été abandonnées.

Dès la phase de fusion des deux bases de données il a été décidé d'abandonner les lignes surnuméraires du jeu de donnée Mobilisation, entraînant la perte acceptable d'un peu plus de 168.000 lignes (8% du total). Cette décision s'appuie sur le fait que le jeu de donnée Incident contenait les dates ainsi que les coordonnées des interventions. Ces variables sont en effet nécessaires à l'analyse d'autres variables, mais également pour mesurer la vélocité des interventions.

À ce premier choix s'ajoute également la décision d'abandonner les lignes d'interventions effectuées par des stations fermées entre 2009 et 2013. En effet, la LFB a fermé à partir de 2013 plusieurs stations dans un soucis d'optimiser ses dépenses. Le choix d'abandonner ces lignes s'est avéré facilité par le faible nombre de lignes

abandonnées (67.450).

L'étape suivante concerne le traitement des valeurs nulles restantes. Tout d'abord, les colonnes *DelayCodeId* et *DelayCodeDescription*, présentant un intérêt potentiel pour l'analyse du temps de réponse, comptait de très nombreuses valeurs nulles (respectivement 30% et 70%). Nous avons donc décidé de remplacer les valeurs nulles de *DelayCodeId* par l'indice précisant l'absence de délais et pour *DelayCodeDescription* par un "None". Cette décision a été prise en raisonnant que si aucune raison n'avait été mentionnée dans le rapport c'est parce que l'équipe d'intervention n'avait pas été délayée.

Enfin, les lignes contenant des valeurs nulles restantes dans les autres variables ont été abandonnées. Elles ne représentaient, après tous les choix précédent, qu'une part très réduite du jeu de donnée disponible (quelques dizaines de millier de lignes, soit moins de 1%).

Après ces nettoyages successifs, le jeu de données compte 1.797.003 entrées. L'analyse des données, et les décisions prises pour le modèle de *machine learning* ont entraîné d'autres choix sur la sélection des lignes que nous évoquerons alors.

1.2.2. Création de variables pour l'analyse du jeu de données

Adresse des stations de la LFB

Avec les premières observations du jeu de données, il est apparu nécessaire d'exploiter les données disponibles et de les enrichir pour permettre des analyses plus complètes. Pour commencer, nous avons décidé d'ajouter au jeu de données les coordonnées géographiques de chacune des stations associées à leurs interventions dans quatre nouvelles colonnes :

- *o_long* et *o_lat* contenant les latitudes et longitudes des casernes.
- *o_x_utm* et *o_y_utm* contenant les coordonnées *Universal Transverse Mercator* (UTM) utilisées ensuite pour calculer les distances

Les adresses des stations ont été collectées *via* le site de la LFB².

L'ensemble des stations considérées dans cette étude, et leurs coordonnées correspondantes, peuvent être consultés à la table en annexe A.1. La Figure 1.1 représente la répartition des stations sur la carte de l'aire urbaine de Londres.

Calcul des distances euclidiennes

Avec les coordonnées des stations disponibles, il a été possible de calculer la distance entre la station et le lieu de l'incident. Mais pour cela il a été également nécessaire de convertir les informations disponibles dans le jeu de don-

²<https://www.london-fire.gov.uk/community/your-borough/>

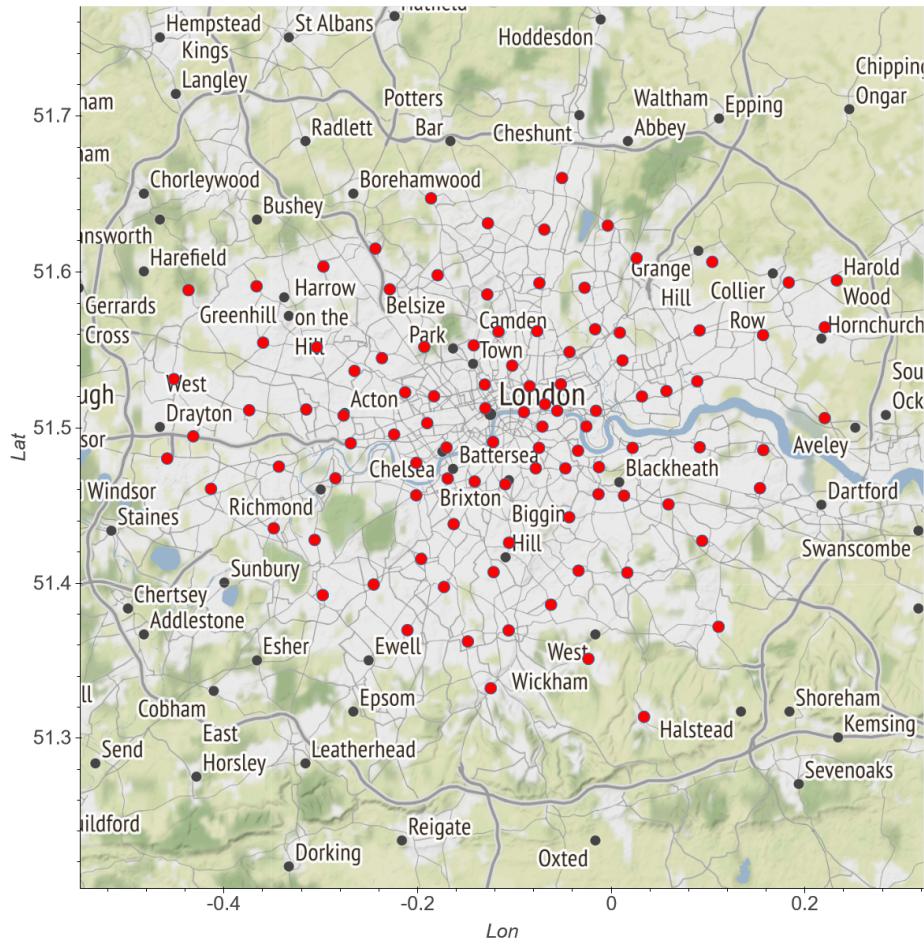


Figure 1.1: London Fire Brigade' Stations

nées. En effet, celui-ci contient les coordonnées des interventions selon un système de coordonnées géographiques cartésiennes. Le seul inconvénient est que le système de référence utilisé est le *Ordnance Survey National Grid*, un référentiel datant de 1936 propre à la Grande-Bretagne qu'il nous faut projeter dans un référentiel universel. Pour les convertir en latitude et longitudes, nous avons eu recours à la bibliothèque *convertbng*³ qui permet les conversions vers des coordonnées géographiques. Elles ont ensuite été stockées pour chaque intervention dans les colonnes *d_long* et *d_lat*.

Comme pour les stations, pour permettre le calcul des distances en mètre il a ensuite été nécessaire de les convertir en coordonnées UTM stockées dans les colonnes *x_utm* et *y_utm*.

Cette conversion a été faite à l'aide de la formule suivante :

$$x_{\text{utm}} = \text{lon} * (k * \frac{\pi}{180}), \quad (1.1)$$

$$y_{\text{utm}} = \log(\tan((90 + \text{lat}) * \frac{\pi}{360})) * k, \quad (1.2)$$

ou $k = 6378137$, soit le rayon en mètres de la Terre.

³<https://pypi.org/project/convertbng/>

Enfin, à partir des colonnes de coordonnées en UTM des stations et des incidents, il a été possible de calculer la distance euclidienne, dans ce cas la distance à vol d'oiseaux, entre ces deux points. Cette valeur a été stockée dans une nouvelle colonne baptisée *dist_euclidian*.

La distance euclidienne a été calculée à partir de la formule suivante :

$$dist_{ij} = \sqrt{(x_{i\text{utm}} - x_{j\text{utm}})^2 + (y_{i\text{utm}} - y_{j\text{utm}})^2}, \quad (1.3)$$

Pour i, j chaque lieu d'incident et station de départ de l'intervention.

Cette formule a été employée également pour calculer la distance relative des stations avec le centre de la ville de Londres. Ces dernières valeurs ont été ensuite discrétisé en 6 quartiles stocké dans une nouvelle colonne exploité dans le cadre des visualisations que nous verrons ensuite.

Analyse des données et visualisations

2.1. Temps d'intervention: analyse et visualisations

Pour l'analyse des temps d'intervention nous nous alignerons sur les pratiques de la LFB qui décompte les temps d'intervention en secondes.

Le temps d'intervention est l'un des principaux objectifs de la LFB. Elle vise ainsi à ce que les premiers secours arrivent dans un temps inférieurs à 360 secondes, et la seconde équipe dans les 480 secondes. Nous allons donc d'abord nous intéresser à la répartition des temps d'intervention. On peut constater dans la figure 2.1 que 50% des unités mobilisées par la LFB arrivent sur les lieux d'intervention en moins de 320 secondes, et que 75% arrivent avant que 430 secondes ne se soient écoulées. Ainsi, la LFB semble remplir les objectifs qu'elle s'est fixé. On notera qu, par soucis de lisibilité, nous avons décidé d'écarter les valeurs extrêmes de cette figure. En effet, les temps d'interventions peuvent ponctuellement atteindre des valeurs extrêmement importantes, sans explications évidentes.

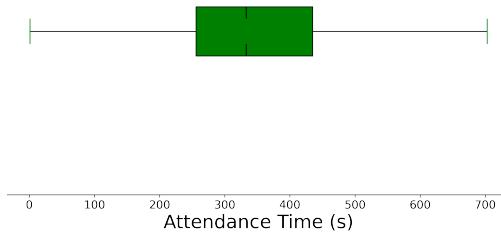
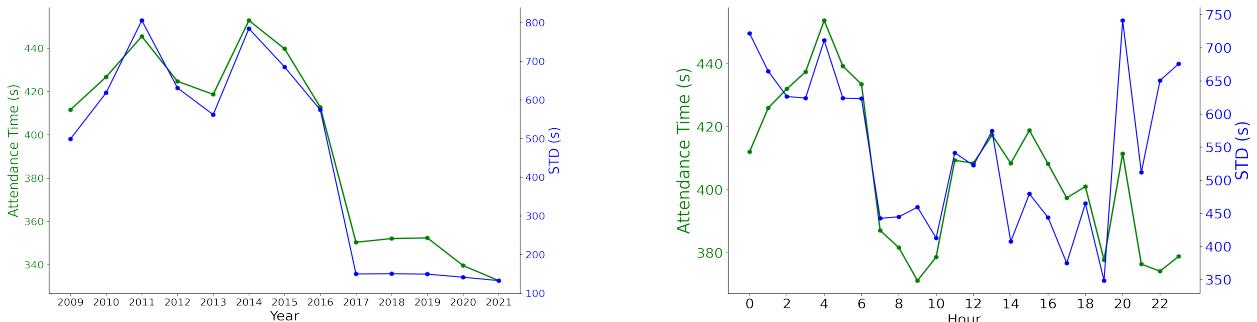


Figure 2.1: étude de la répartition des temps d'intervention

Il apparaît pertinent de s'intéresser à l'évolution du temps d'intervention en fonction des années et des heures. Nous pouvons constater dans un premier temps une évolution marquée du temps d'intervention moyen entre 2013 et 2016 suivi par une forte diminution de ce dernier(2.2a). Il est très probable qu'il s'agisse d'une conséquence de la réorganisation de la LFB suite à la réforme engagée en 2013. Lors de cette réforme, la LFB a fermé plusieurs stations peu actives de la métropole londonienne, afin de mieux allouer les moyens restants. Après une hausse due probablement au temps d'adaptation sur cette période, le temps moyen d'intervention a connu une très forte réduction, suivie d'une stabilisation.

Un test statistique effectué sur les temps d'intervention par années confirme l'impact de cette réforme. En effet, l'écart type est très fortement et positivement corrélée à la moyenne, mais toutes deux chutent fortement en 2017



(a) Temps d'intervention moyen(vert) et son écart type(bleu) par années

(b) Temps d'intervention moyen (vert) et son écart type(bleu) par heures

Figure 2.2: évolution du temps d'intervention moyen et son écart type

et deviennent stables. On peut donc confirmer que les données antérieures à 2016 ne seraient pas représentatives des délais d'intervention actuels.

Les temps d'intervention moyens par heures sont marqués par deux pics autour de 4 heure du matin et entre 11h et 20h(2.2b). Le premier peut s'expliquer par le temps nécessaires aux équipes de secours pour se mobiliser au milieu de la nuit. Celui entre 10h et 20h quant à lui semble plus logiquement correspondre à une suractivité des unités de secours entre ces heures, comme nous pouvons le voir dans la figure 2.3, mais aussi en raison peut-être de l'augmentation du trafic routier.

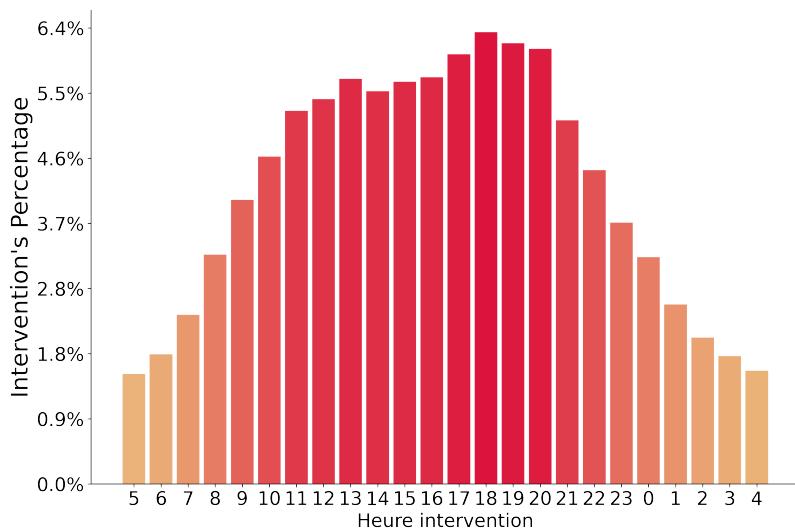


Figure 2.3: Part des interventions par heures de la journée

De fait, un test statistique effectué sur les temps d'intervention par heures laisse à penser qu'il y a une très grande variabilité dans le temps des interventions. En effet, le coefficient de corrélation entre la moyenne et son écart-type par heures est $\rho = 0.513$, $P < .05$. D'après les seuils proposés par Cohen (1988)¹, cette corrélation est moyenne et positive.

¹Cohen, J. 1988, *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates.

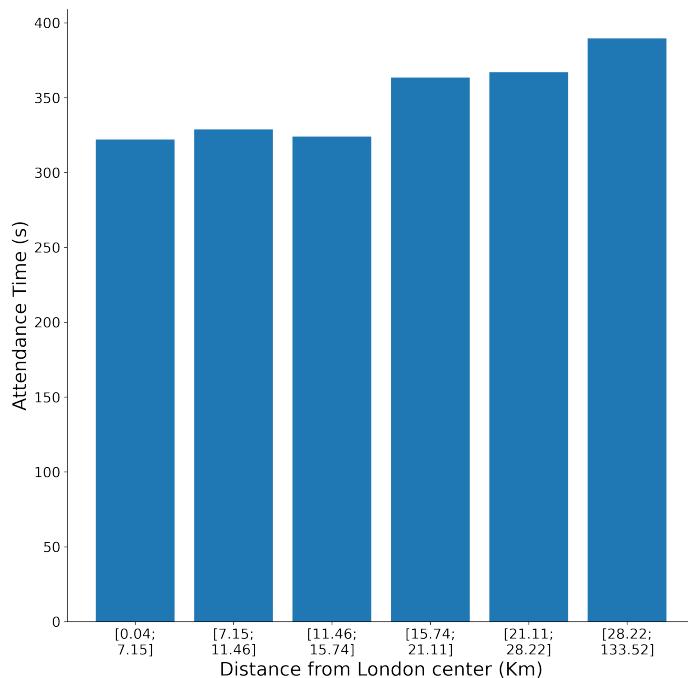


Figure 2.4: Temps d'intervention moyen en fonction de la distance relative avec le centre de Londres après 2016

Enfin, on peut s'intéresser au rapport entre la distance avec le centre de l'aire urbaine de Londres et le temps d'interventions. En s'appuyant sur les résultats des tests statistiques et pour une meilleure représentativité, on s'intéresse seulement aux interventions post-2016. On peut constater ainsi dans la figure 2.4 que les stations du centre ont un temps moyen d'intervention inférieur aux 360 secondes, objectif fixé par la LFB. Au delà de 15 km du centre de Londres on peut observer une augmentation du temps moyen d'intervention, qui se maintient toutefois en dessous de 360 secondes, sauf pour le dernier quartiles. Cependant, il s'agit de stations éloignées de plus de 30 km du centre, et dans des zones plus rurales, ce qui peut expliquer cette variation. Cette évolution du temps interroge donc sur les distances parcourues par les stations.

2.2. Distances d'intervention: analyse et visualisations

Comme on l'a vu précédemment la distance relative avec le centre de Londres a une influence assez marquée sur le temps moyen d'intervention. Il apparaît donc pertinent de s'intéresser à la variable distance en elle-même, notamment dans le cadre de la préparation de l'entraînement d'un algorithme pour la prédictions du temps d'intervention.

Dans la figure 2.5 on constate tout d'abord que les stations du centre de l'aire urbaine de Londres concentrent la majorité des interventions. La densité du tissu urbain peut expliquer cette variation.

la figure 2.6 nous permet de visualiser que, à l'inverse de la figure précédentes, les stations de la périphéries de la zone urbaine londonienne accumulent des distances moyennes parcourues plus grandes que celle du centre. Ici, la densité du réseau des stations, plus relâchées dans les périphéries, apparaît comme une explication logique de cette variation.

Enfin, comme pour le temps de réponse moyen, on peut observer (2.7a) l'impact de la réforme amorcée en 2013, avec pareillement une régularisation après 2016. Mais également l'impact (2.7b) de la suractivité entre 10h et 21h observé précédemment (2.3).

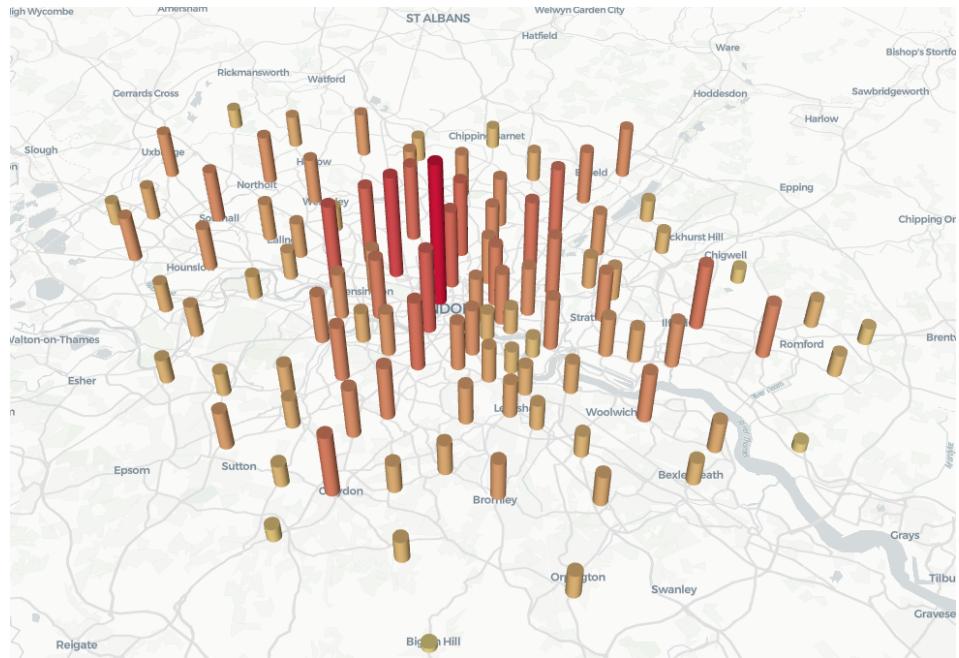


Figure 2.5: Interventions par station

Toutefois, comme nous l'avons précisé lors de l'étude du jeu de données, les distances étudiées ici sont les distances euclidiennes séparant les stations des lieux d'incidents. Elles permettent d'étudier très rapidement les limites des espaces attribuées aux unités d'intervention. Mais pour un calcul plus précis des temps d'interventions à prédire à l'aide d'un *machine learning*, le calcul des distances réelles, prenant en compte le trajet *via* les voies de circulation, semble plus pertinent.

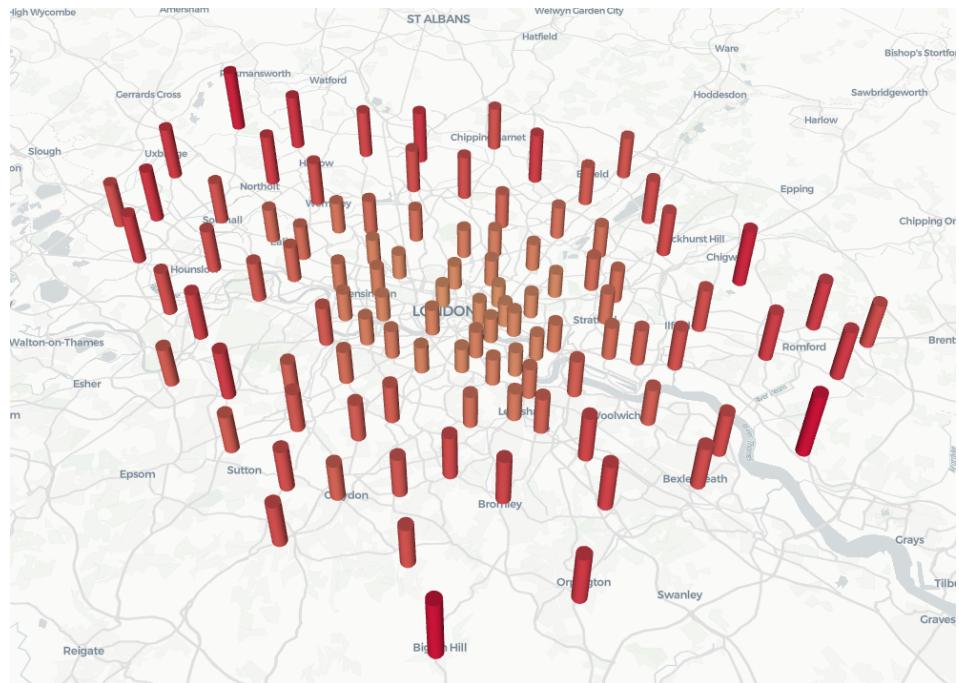
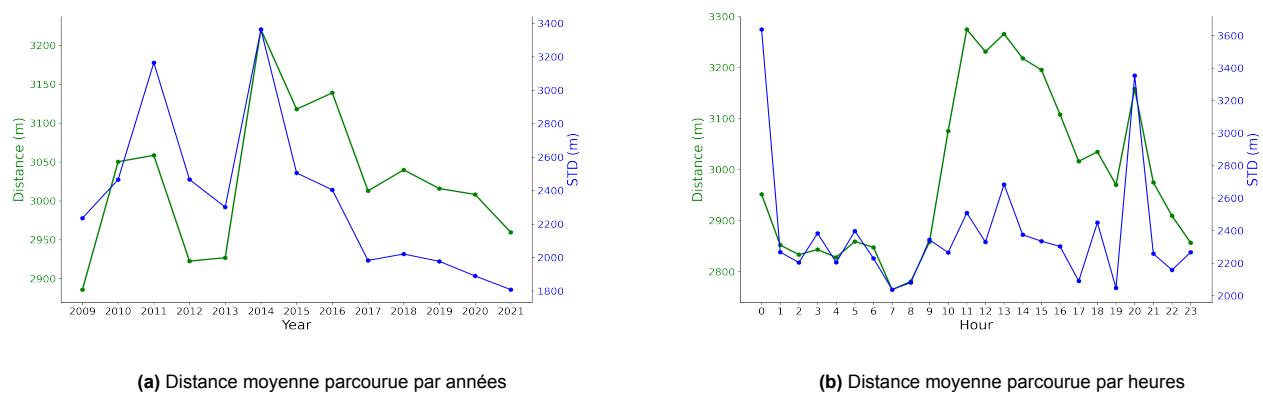


Figure 2.6: Carte des distances moyennes par station

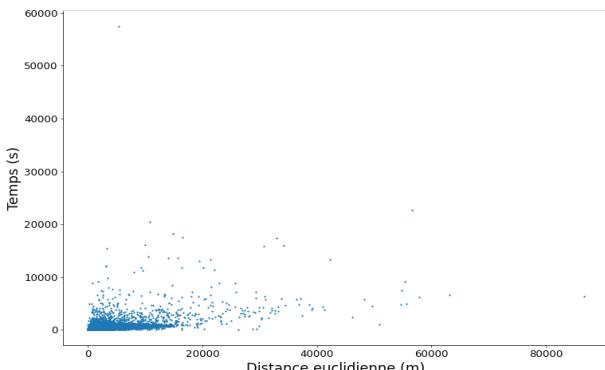


Classification et Prédition

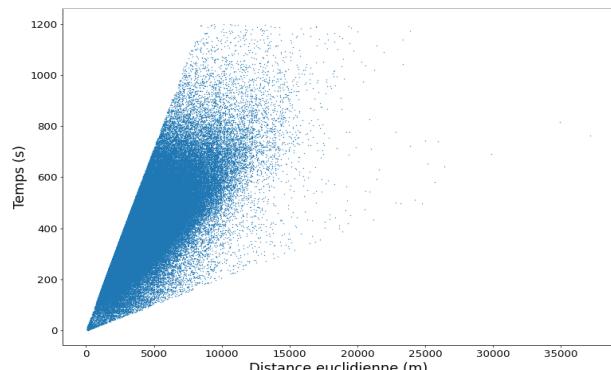
3.1. Échantillonage

Dans cette analyse, on s'intéresse à la prédition des temps d'interventions des Pompiers londoniens. L'information la plus pertinente pour déduire ce temps d'intervention, est la distance à parcourir par les unités de la LFB. Si l'on regarde la distribution initiale des temps d'intervention en fonction de la distance (Figure 3.1a), on peut remarquer que certaines données pourraient ne pas refléter les temps que peuvent mettre les pompiers en action. On peut donc constater que dans les données doivent encore se trouver quelques anomalies :

- Des interventions très proches de la station, mais pour lesquelles les pompiers ont pris plusieurs minutes
- Des interventions lointaines, mais pour lesquelles les pompiers ont pu recourir à des véhicules plus rapides¹
- Des interventions différées (relèves, exercices, effort prolongé de lutte contre le feu...)



(a) Avant échantillonage



(b) Après échantillonage

Figure 3.1: Distribution des temps de réponse en fonction de la distance

Comme nous souhaitons nous focaliser sur les évènements nécessitant une intervention classique (véhicules), on peut appliquer quelques filtres supplémentaires :

- Temps d'intervention supérieur à 0 seconde

¹Hélicoptères, formule 1, téléporteur, ou encore Sleipnir, fils de Loki, Destrier d'Odin, le cheval à 8 jambes capable de survoler les mers

- Une distance euclidienne supérieure à 100m
- Ne garder que les données enregistrées après 2016 (voir Figure 2.2a)
- Ne conserver que les données dont la vitesse calculée est comprise entre 7 m.s^{-1} (25 km.h^{-1}) et 50 m.s^{-1} (180 km.h^{-1})

Ces filtres sont très restrictifs, car ils réduisent la base de données de 78.58%. La distribution du temps en fonction de la distance de cet échantillon est rapporté Figure 3.1b.

La distribution des données ne semble pas s'offrir à l'utilisation d'une régression linéaire. Nous suggérons qu'elles dépendent de trop de facteurs différents pour offrir une prédiction satisfaisante sans davantage de détails logistiques (absent de la base de données). Nous avons alors imaginé de produire des informations pertinentes et descriptives, en fonction d'un lieu d'intervention. Il nous faut donc déterminer pour chaque point de la carte, la station qui devrait intervenir

3.2. *Pre-processing* et Entraînement

La distribution des incidents en fonction de leurs coordonnées, en fonction de la station qui est intervenue, semble convenir à une classification par la méthode des K moyens. Cependant, en observant cette distribution (Figure 3.2a), nous nous sommes interrogés sur un moyen d'améliorer la qualité d'un apprentissage par les K moyens. En effet, les territoires sur lesquels les stations ont été amenées à intervenir se chevauchent.

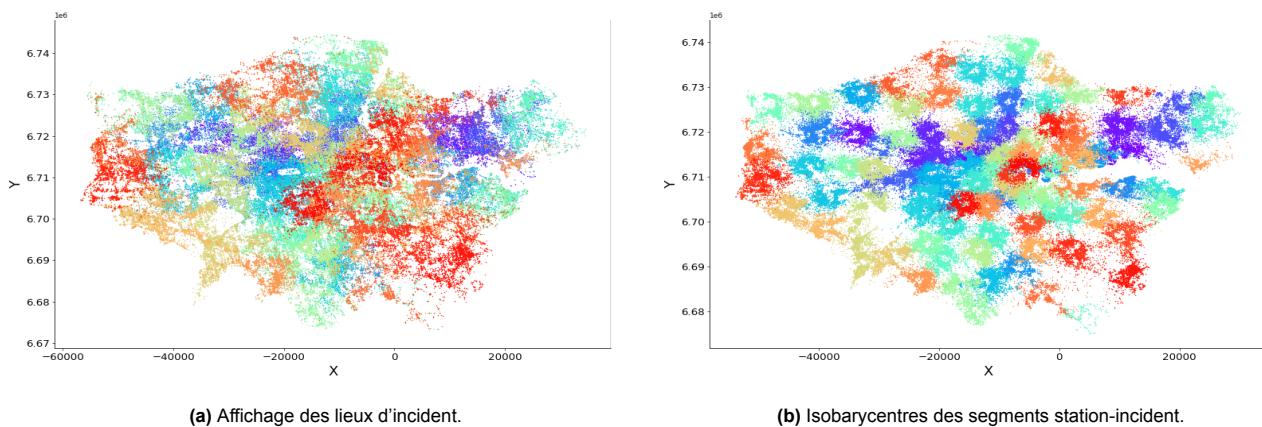


Figure 3.2: Représentation géographique des incidents coloré par station ayant participé à l'intervention.

Une solution serait d'entraîner le modèle, avec un lieu d'incident tronqué, représentant un fragment de la distance station-incident. Un exemple de cette troncature est affiché Figure 3.2b, dans laquelle nous avons représenté les isobarycentres des segments station-incident, coloré par station.

Pour entraîner le modèle nous avons sélectionné aléatoirement 80% des données pour l'entraînement, le reste pour le test. Le modèle des K moyens est initialisé avec autant de centres qu'il y a de stations (102). Nous

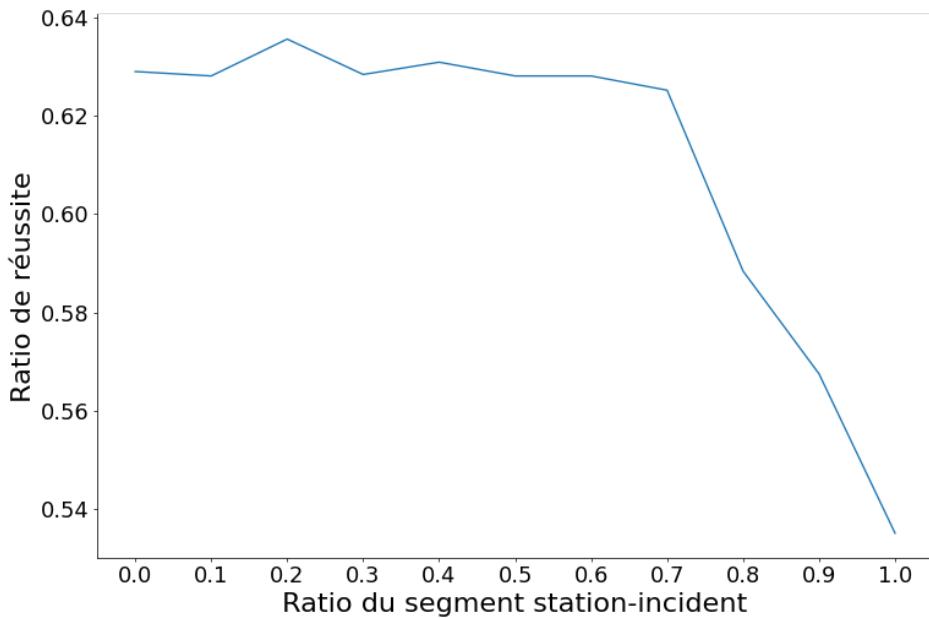


Figure 3.3: Résultats de la classification par les K moyens, en fonction du fragment du segment utilisé pendant l'entraînement.

avons spécifié dans ce modèle que les coordonnées des centres correspondent aux coordonnées des stations. La tâche du modèle consiste à indiquer pour chaque coordonnées d'incident, à quelle station il appartient. Lors de l'entraînement, nous avons utilisé les coordonnées correspondantes aux fragments des segments station-incident. Lors de la phase de test, nous avons utilisé les coordonnées des incidents uniquement.

Lorsqu'on observe les performances du modèle des K moyens (Figure 3.3), on peut souligner que d'utiliser les fragments de segment a bien amélioré la classification. Cependant, la performance avec la méthode K moyens n'est pas fantastique (max 63%). Nous pensons que la cause principale de ce score est le fait que les stations ont fréquemment pu intervenir sur des secteurs communs. Une solution pertinente serait, pour chaque lieu d'incident, de déterminer quels sont les stations qui sont susceptibles d'y intervenir.

Bien sûr, on pourrait simplement utiliser la distance, mais il serait difficile de fixer le seuil : un incident en campagne pourrait faire intervenir 2-3 stations situées à 10-15 km, mais un incident en ville pourrait faire intervenir 4-5 stations situées à 3 km. Gérer ces spécificités individuellement pourrait s'avérer plus compliqué qu'appliquer une méthode de ML. Cependant, les K moyens ne renvoie pas de probabilité associée à la classe. Nous choisissons donc d'utiliser les Mixtures Gaussiennes (*Gaussian Mixture Model*, GMM), une généralisation de l'algorithme des K moyens, mais qui leur associe une covariance. Cette covariance peut capturer les caractéristiques de chaque station, en fonction de leur historique d'intervention.

On pourra alors, pour chaque lieu d'incident, déterminer quelles stations sont susceptibles d'intervenir. On pourra également vérifier si parmi les stations candidates, la station du jeu de données est bien présente.

Nous avons donc répliqué la méthode des K moyens en utilisant des Mixtures Gaussiennes. Le modèle des Mixtures Gaussiennes est initialisé avec autant de centres qu'il y a de stations (102). Nous avons spécifié

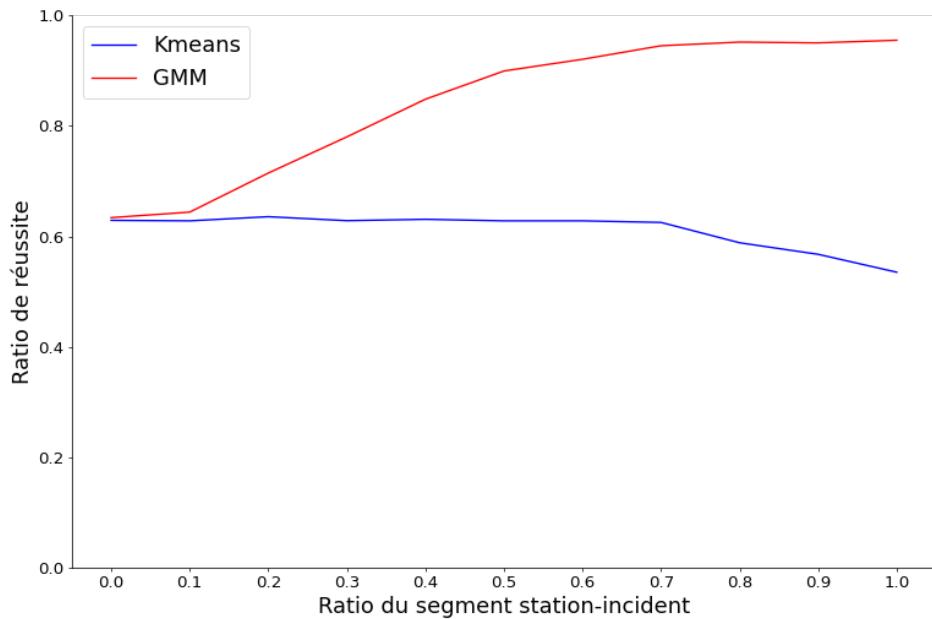


Figure 3.4: Comparaison des performances GMM, K moyens

dans ce modèle que les coordonnées des centres correspondent aux coordonnées des stations, ainsi que le ratio d'intervention de chaque stations. Ce dernier paramètre agit comme une "gravité", donnant davantage d'importance aux centres dont le taux d'intervention est élevé. La tâche du modèle consiste à indiquer pour chaque coordonnées d'incident, à quelle station il appartient. Pour mesurer l'efficacité du modèle, nous avons utilisé les probabilité des prédictions. Ainsi, pour chaque incident, nous avons contrôlé si parmi les stations dont la probabilité d'appartenance est non-nulle (supérieure à 0.1 %), se trouve la station ayant effectivement réalisé l'intervention. Nous avons ainsi pu comparer les performances des K moyens et du GMM, Figure 3.4.

Sans surprise, le GMM est plus performant que le K moyens, car il est moins restrictif et nous permet d'analyser la probabilité d'appartenance d'un incident aux stations environnantes. Comme ce modèle est plus souple, il est également plus performant avec les données brutes (lieu de l'incident), car il capture les caractéristiques relatives à la distribution géographique des incidents, qui peut varier entre les stations. Une représentation de ces GMM et de leurs covariance peut être visualisée Figure 3.5.

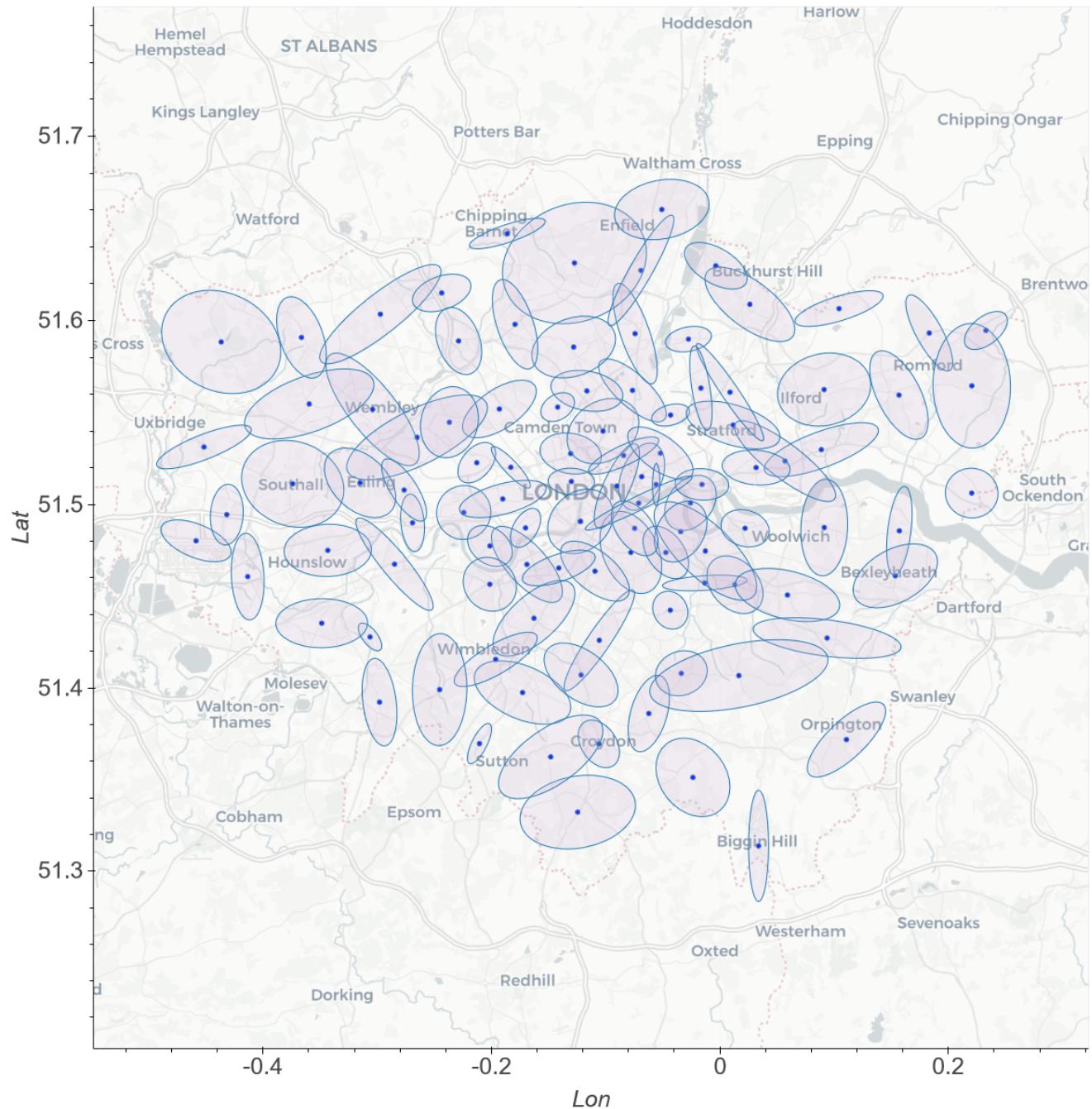


Figure 3.5: Projection géographique des GMM et de leurs covariances

Calcul des itinéraires

4.1. Différentes procédures possibles

Plusieurs approches sont possibles pour le calcul d'itinéraire. Nous avons comparé les différents types de formats de données géographiques, ainsi que leurs performances d'utilisation. D'autres optimisations sont possibles, plus particulièrement au niveau des algorithmes de recherche. Cependant, au vu de la quantité de données, nous avons préféré rechercher une méthode optimale pour le calcul d'itinéraires sur une grande quantité de données. Une illustration des différents résultats obtenus pour la recherche d'itinéraire a été rapporté Figure 4.1. Pour simplifier les comparaisons, nous avons décrit les méthodes optimales pour chaque type de fichiers géographiques, que l'on peut trouver sur le site OpenStreetMap¹.

4.1.1. Extension OSM

Les fichiers OSM peuvent directement être utilisés avec OSMNX, un utilitaire de représentations de *graphs* en python.

La méthode avec OMSNX est, comme son nom l'indique, une méthode qui couple les réseaux de type OSM (i.e., Open Street Map) et les méthodes de la librairie NetworkX.

C'est la méthode la plus fiable pour calculer des itinéraires, car les réseaux au format OSM sont complets et contiennent les coordonnées des noeuds (WGS84), les liens qui relient ces noeuds, ainsi que la nature de la relation entre ces noeuds.

Si on traite les données à partir du fichier OSM (i.e., *standalone*), bien que le taux de compression des fichiers OSM soit important, une fois décompressé la mémoire requise pour travailler sur le réseau est extrêmement importante. Pour illustrer, le fichier OSM qui contient les données du grand Londres compressé pèse 1.6 Go, et plus de 18 Go décompressé.

Si l'on veut traiter les itinéraires un par un (i.e., dynamique), il est possible de formuler des requêtes directes pour récupérer l'itinéraire entre deux points. Ce système de requête point-à-point est gracieusement mis à disposition

¹<https://www.openstreetmap.org/>

par OpenStreetMap, mais l'hébergement est payant et donc le nombre de requêtes est, à juste titre, limité.

Ce mode présente donc deux difficultés d'utilisation : en *standalone*, les caractéristiques machine requises sont extrêmement importantes; en dynamique, le nombre de requêtes est limité, cette approche n'a donc pas été retenue pour traiter nos quelques 1.7M de données.

4.1.2. Extension Shapefile

Les données au format Shapefile résument les informations de relation entre différents points sur une carte données, cependant ils manquent d'informations sur la topologie. C'est également une représentation simplifiée de la géographie, le calcul d'itinéraire par cette approche peut comporter des erreurs d'approximation. Cependant, le temps de calcul en utilisant cette approche est fortement réduit, ce qui laisse à penser que dans le cas du calcul d'itinéraires longues distances, elle est optimale.

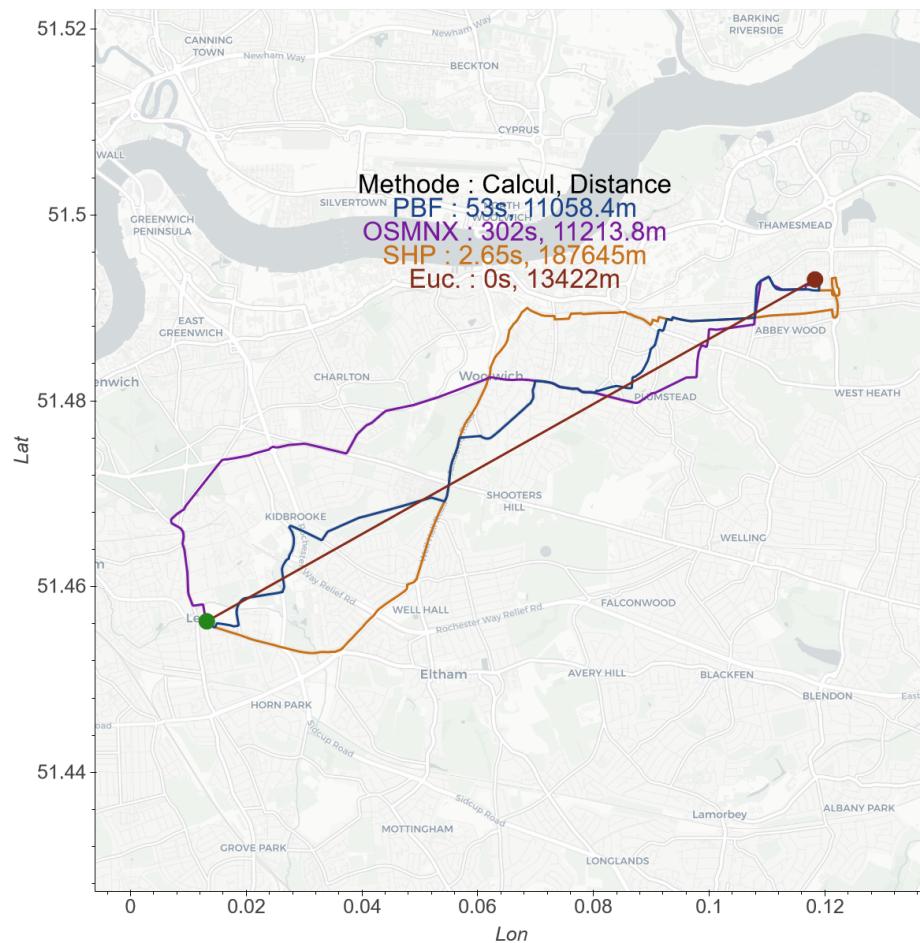


Figure 4.1: Illustration des différentes méthodes de calcul d'itinéraire, avec le trajet calculé, la distance totale du trajet, et le temps de calcul. La distance euclidienne est plus importante que les trajets car elle est calculée sur la projection UTM des coordonnées, tandis que les trajets calculés avec OSMNX et PBF calcul sur les coordonnées avant de réaliser la projection, et accumulent donc moins d'erreur.

4.1.3. Extension PBF

Les fichiers PBF sont des versions simplifiées des fichiers OSM. Le fichier à importer est réduit en taille, et la librairie Pyrosm permet de sélectionner certaines caractéristiques de la carte (e.g., voies routières, voies pédestres, bâtiments...) avant de les convertir en réseau, pour être utilisé avec NetworkX. La librairie Pyrosm offre également la possibilité de charger des cartes en fonction d'un périmètre. Cette approche est très intéressante lorsqu'il s'agit d'optimiser un calcul distribué pour l'extraction de coordonnées en masse, comme c'est le cas pour nos 1.7M de données.

4.2. Application

Nous avons constaté que le calcul PBF présente un temps d'exécution presque 6 fois inférieur au calcul avec OSMNX directement, avec une distance calculée très proche. Si la méthode PBF semble être à privilégier pour l'estimation des distances de trajets sur une grande base de données, un temps d'exécution total de 50 secondes diminue fortement l'intérêt de cette méthode pour une application en temps réel.

En revanche, la méthode Shapefile est extrêmement rapide, et il est à noter qu'elle comprend le chargement initial des données qui, au contraire des deux autres méthodes, contient toute la zone géographique concernée (i.e., l'aire urbaine de Londres). Ainsi, mis à part le chargement initial de 2 secondes, le temps de calcul des itinéraires individuels est très faible.

Cependant, la distance du trajet estimé est très grande, signe que la méthode de calcul *via* la méthode Shapefile est source de surestimations.

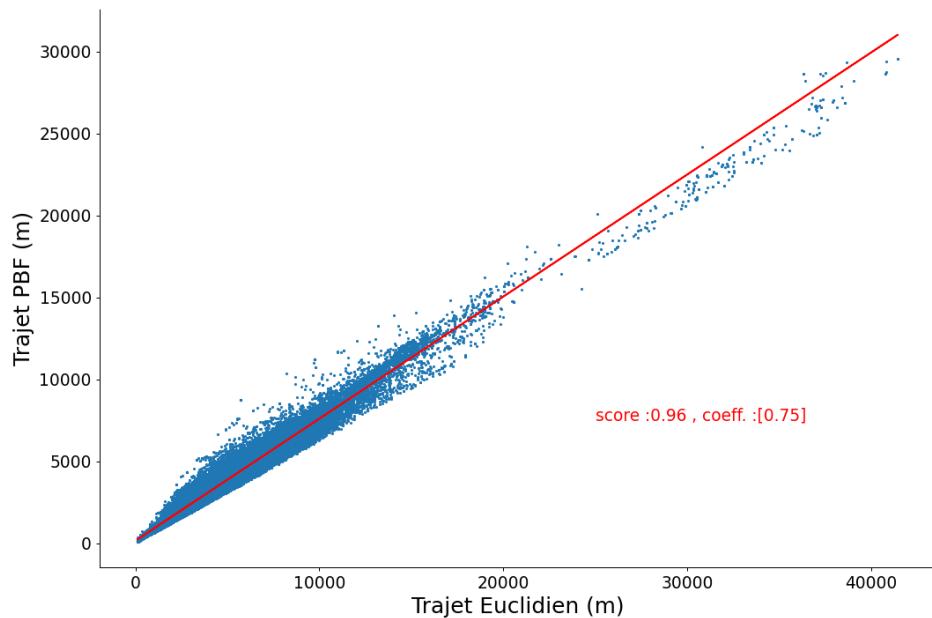


Figure 4.2: Relation entre distances euclidiennes et distances calculées avec PBF

L'idéal serait donc de conjuguer un travail préalable de calcul des trajets en PBF, et d'utiliser ces données en déduisant la distance de trajet probable par le calcul de la distance euclidienne. Nous avons donc analysé la relation entre les distances calculées en PBF et en distance euclidienne, rapportée Figure 4.2.

Nous voyons ainsi qu'il est possible d'estimer une distance très proche de la distance calculée avec PBF, en appliquant un coefficient de 0.75 à la distance euclidienne.

4.3. Algorithme

La procédure que nous avons sélectionné pour la prédiction des temps de réponse des pompiers londoniens est donc :

- Sélection d'un lieu d'incident sur la carte
- Grâce à un modèle pré-entraîné de GMM, déterminer quelles stations sont susceptibles d'intervenir
- Pour chaque station sélectionnée, calculer la distance euclidienne qui la sépare du lieu de l'incident
- Appliquer le coefficient de conversion pour déduire une valeur proche de la distance réelle de conduite
- Pour chaque station sélectionnée, prédire à l'aide d'un modèle linéaire pré-entraîné correspondant à la station, le temps d'intervention en fonction de la distance de conduite estimée
- Afficher les informations sur la carte et les résultats du test linéaire de chaque station
- Ajouter une estimation du trajet grâce aux données Shapefile pour le côté esthétique

Cette procédure nous permet de produire un résultat global :

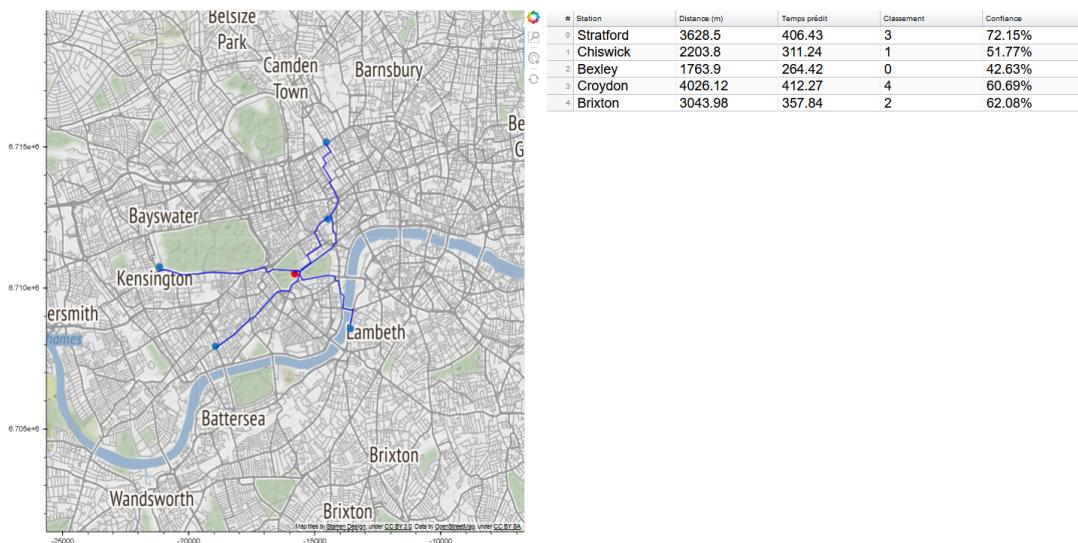


Figure 4.3: Résultats final.

Conclusion

L'analyse des temps d'intervention confirme que la LFB atteint ses objectifs fixé, avec un temps moyen inférieur à 330 secondes lissé sur toute la région. Comme on l'a vu également, seules les stations les plus éloignées du centre de Londres ne parviennent pas à intervenir en dessous de 360 secondes en moyenne, mais restent dans une marge acceptable (aux alentours de 370 secondes).

Les événements sont concentrées dans la zone centrale de l'aire urbaine de Londres. les stations de cette zones parcourrent des distances moyennes relativement plus faibles que celle des périphéries, une variation expliquée par la dispersion plus importante des stations plus on s'éloigne du centre.

Un modèle de prédiction entraîné directement sur les données pourrait ne pas refléter les nombreuses particularités que peut présenter l'intervention des pompiers londoniens. En revanche, l'attribution automatisée de station pouvant potentiellement intervenir sur un point, permet d'affiner l'analyse des statistique d'une station et d'en prédire un temps d'intervention. Cette méthode semble également moins exigeante en terme de ressources computationnelles, et permet d'obtenir très rapidement des résultats tangibles.

Pour affiner les analyses et prédictions il serait intéressant de les croiser avec les données de circulations dans l'aire urbaine de Londres. Nous n'avons pas pu collecter d'informations suffisamment complètes sur ce sujet, les données libres disponibles ne couvrant généralement que les jours ouvrés et aux horaires de travail.

A

Annexe

Table A.1: Station's name and coordinates used in this study.

Index	Station	lat	lon	Index	Station	lat	lon
37	Acton	51.507845	-0.276483	91	Lambeth	51.490753	-0.122097
62	Addington	51.351158	-0.023605	92	Lambeth River	51.490753	-0.122097
0	Barking	51.529820	0.088809	70	Lee Green	51.456260	0.013175
27	Barnet	51.647239	-0.186118	74	Lewisham	51.457223	-0.013226
99	Battersea	51.467349	-0.168972	24	Leyton	51.563270	-0.016630
58	Beckenham	51.407975	-0.033810	25	Leytonstone	51.561057	0.008853
18	Bethnal Green	51.527910	-0.052074	30	Mill Hill	51.614985	-0.243591
55	Bexley	51.461121	0.153563	19	Millwall	51.500806	-0.025725
59	Biggin Hill	51.313730	0.033850	94	Mitcham	51.397482	-0.172787
89	Brixton	51.463652	-0.109389	75	New Cross	51.473858	-0.047339
60	Bromley	51.406732	0.016589	88	New Malden	51.399077	-0.245362
83	Chelsea	51.487233	-0.170153	64	Norbury	51.407077	-0.121559
23	Chingford	51.629740	-0.003590	85	North Kensington	51.522718	-0.212865
52	Chiswick	51.489993	-0.269113	39	Northolt	51.554662	-0.359569
90	Clapham	51.465400	-0.141051	77	Old Kent Road	51.486999	-0.074664
63	Croydon	51.369470	-0.106000	61	Orpington	51.371835	0.110792
1	Dagenham	51.559507	0.156780	101	Paddington	51.520219	-0.183039
72	Deptford	51.485205	-0.034350	31	Park Royal	51.536515	-0.265157
76	Dockhead	51.500708	-0.071144	78	Peckham	51.473869	-0.077979
2	Dowgate	51.510003	-0.090206	13	Plaistow	51.520068	0.031699
38	Ealing	51.511740	-0.315039	71	Plumstead	51.487420	0.091370
67	East Greenwich	51.487034	0.022049	20	Poplar	51.510880	-0.015706
12	East Ham	51.523548	0.056981	65	Purley	51.332205	-0.124477
42	Edmonton	51.627217	-0.069120	96	Richmond	51.467465	-0.284858
68	Eltham	51.450660	0.059100	8	Romford	51.593238	0.183295
41	Enfield	51.660277	-0.050856	51	Ruislip	51.588373	-0.436624
56	Erith	51.485616	0.157079	21	Shadwell	51.510868	-0.055828
35	Euston	51.527723	-0.130668	3	Shoreditch	51.526625	-0.084321
53	Feltham	51.46070	-0.413420	57	Sidcup	51.427203	0.093850
28	Finchley	51.597949	-0.179350	102	Soho	51.512489	-0.130081
73	Forest Hill	51.442390	-0.043444	40	Southall	51.511223	-0.373985
82	Fulham	51.477418	-0.201330	43	Southgate	51.631254	-0.127368
69	Greenwich	51.474642	-0.012700	47	Stanmore	51.603466	-0.297278
15	Hainault	51.606490	0.104310	4	Stoke Newington	51.562056	-0.076530
81	Hammersmith	51.495618	-0.224269	14	Stratford	51.543210	0.011631
6	Harold Hill	51.594600	0.232900	87	Surbiton	51.392245	-0.297993
46	Harrow	51.590833	-0.366337	79	Sutton	51.369629	-0.210518
48	Hayes	51.494466	-0.431728	100	Tooting	51.437991	-0.162768
49	Heathrow	51.480183	-0.458630	45	Tottenham	51.592905	-0.074228
29	Hendon	51.588901	-0.228776	97	Twickenham	51.435297	-0.348598
54	Heston	51.474977	-0.343292	80	Wallington	51.362365	-0.148124
50	Hillingdon	51.531257	-0.451701	26	Walthamstow	51.589890	-0.027524
10	Holloway	51.561766	-0.116461	98	Wandsworth	51.456521	-0.201461
5	Homerton	51.548634	-0.043184	32	Wembley	51.551733	-0.304071
7	Hornchurch	51.564543	0.220518	9	Wennington	51.506171	0.220432
44	Hornsey	51.585610	-0.128019	36	West Hampstead	51.551940	-0.193074
17	Ilford	51.562493	0.091230	93	West Norwood	51.425983	-0.105596
11	Islington	51.539912	-0.102337	22	Whitechapel	51.515125	-0.068433
84	Kensington	51.502965	-0.190099	33	Willesden	51.544717	-0.236838
34	Kentish Town	51.552957	-0.142053	95	Wimbledon	51.415562	-0.196295
86	Kingston	51.427797	-0.306333	16	Woodford	51.608830	0.026243
:	:	:	:	66	Woodside	51.386048	-0.062230