Audit comparatif des données Open Data des arrêts de bus de la région Île-de-France - mise à jour et conclusion de mai 2019

Publié le 24 mai 2019 par Jungle Bus (https://junglebus.io)

Ce document conclut une analyse qui a débutée avec l'audit publié en avril 2018 (http://junglebus.io/iledefrance/audit_2018_04/) et qui a été suivie de 3 mises à jour successives (juillet 2018 (http://junglebus.io/iledefrance/audit_2018_07/), novembre 2018 (http://junglebus.io/iledefrance/audit_2018_11/) et février 2019 (http://junglebus.io/iledefrance/audit_2019_02/)).

Introduction

Île-de-France Mobilités publie en Open Data les données des lignes et des arrêts de transport d'Île-de-France.

La communauté OpenStreetMap cartographie le monde entier rue après rue et participe à la création de la plus grande base de données géographique sous licence libre. On y retrouve donc naturellement des objets de transports en commun.

Une première étude a été réalisée par Jungle Bus en avril 2018 pour comparer les données d'Île-de-France Mobilités et celles d'OpenStreetMap sur la région Île-de-France afin d'évaluer dans quelle mesure ces jeux de données pouvaient s'enrichir mutuellement. Une méthodologie ainsi que des indicateurs y ont été introduits ; le présent document propose d'étudier les évolutions de ces indicateurs suite aux mises à jour de données effectuées depuis, aussi bien dans les données officielles que dans OpenStreetMap. À noter que trois précédentes mises à jour ont déjà été publiées dans ce but.

L'audit mené l'année dernière a révélé des disparités de modélisation des objets de transport et propose en conséquence de comparer le plus petit dénominateur commun entre les deux sources : c'est le *routepoint*, défini comme un arrêt desservi par une ligne vers une destination.

Sur cette base, l'étude a mis en lumière l'hétérogénéité des contributions OpenStreetMap ainsi que leur non-exhaustivité, notamment concernant les lignes de bus.
Malgré cette incomplétude, sur une sélection limitée permettant de garantir une comparaison des mêmes objets entre les deux référentiels, on constate qu'OpenStreetMap possède des données d'excellente qualité qui permettent de proposer des enrichissements aux données officielles d'Île-de-France Mobilités sur un panel de sujets utiles à l'information voyageur.

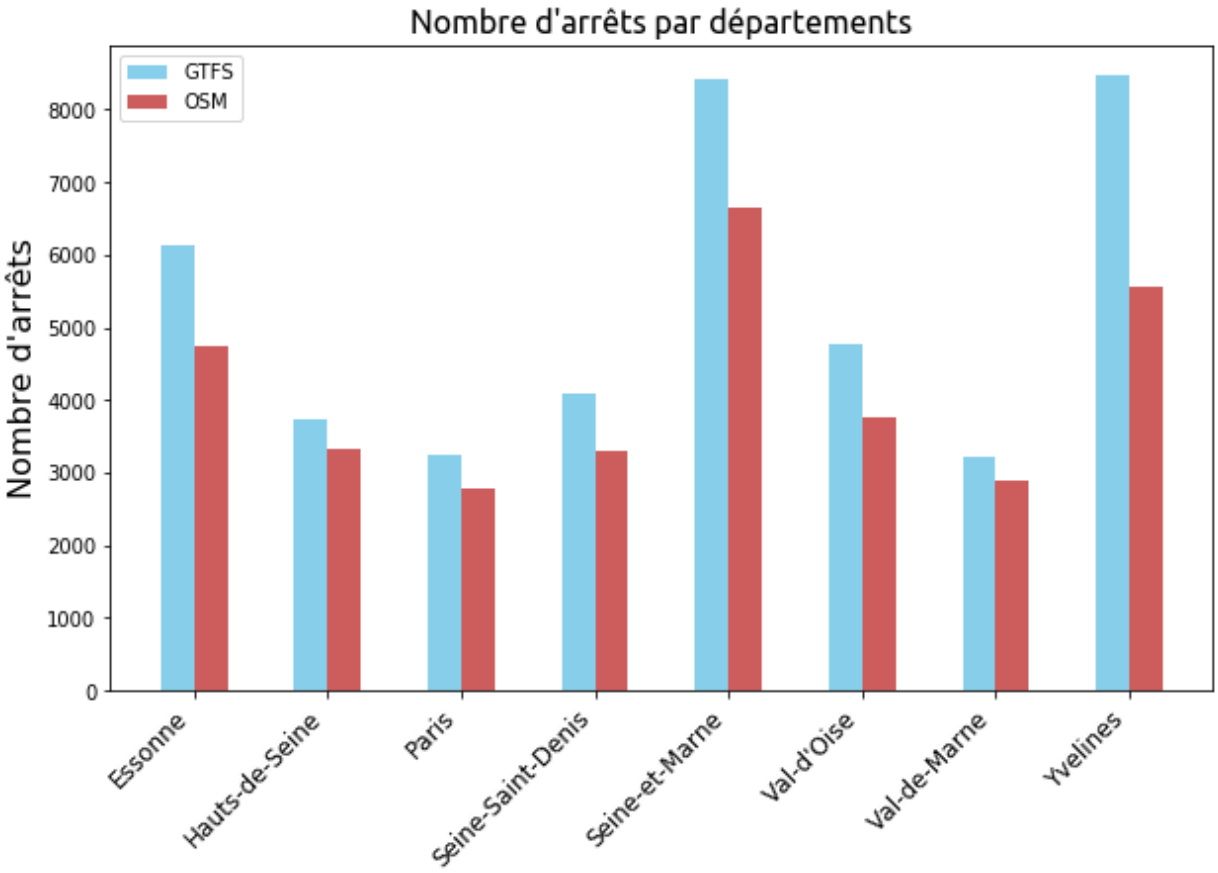
Vous pouvez consulter l'audit initial (http://junglebus.io/iledefrance/audit_2018_04/) dans son intégralité, ainsi que ses mises à jour de juillet 2018 (http://junglebus.io/iledefrance/audit_2018_07/), novembre 2018 (http://junglebus.io/iledefrance/audit_2018_11/) et et février 2019 (http://junglebus.io/iledefrance/audit_2019_02/) pour obtenir la méthodologie de calcul de chaque indicateur proposé ainsi que des conclusions plus détaillées.

Voyons à présent de quelle manière les chiffres ont évolué.

Analyses quantitatives

Nombre d'arrêts

Voici la répartition du nombre relatif d'arrêts dans les deux référentiels, synthétisée par département.



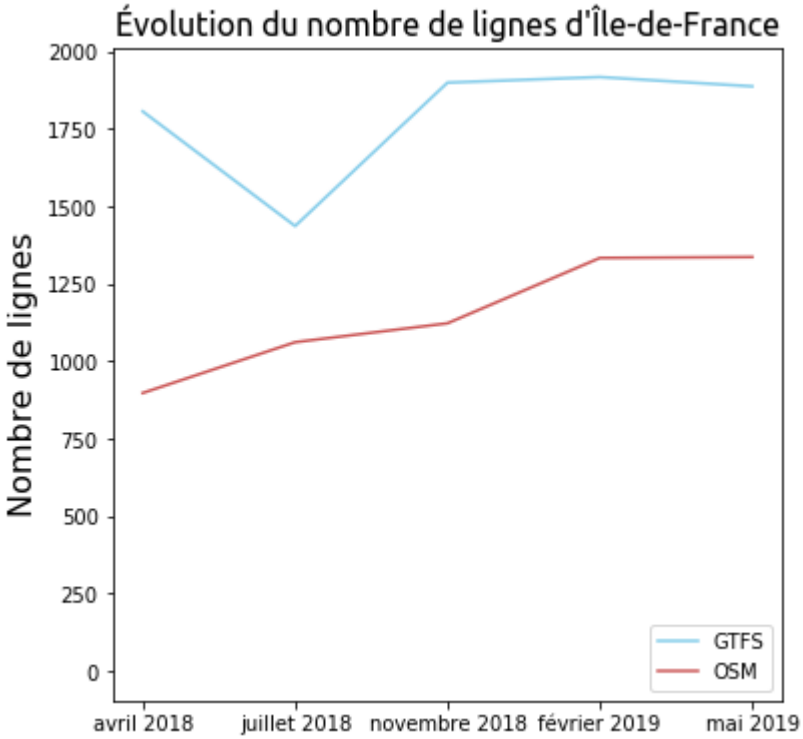
Le pourcentage global de couverture d'OpenStreetMap en arrêts sur la région Île-de-France est de **78 %**.

Comme lors de nos précédentes mises à jour, on constate une très légère progression de ce chiffre.

Nombre de lignes

Le pourcentage global de couverture en lignes de bus sur la région Île-de-France est de **71 %**.

Voici un graphique de l'évolution du nombre de lignes dans les données d'Île-de-France Mobilités et dans OpenStreetMap au cours de nos différents audits :



Le décrochage dans les données Open Data officielles en juillet 2018 correspond aux vacances d'été, durant lesquelles de nombreuses lignes de transport scolaire n'étaient plus présentes dans les données d'offre publiées par Île-de-France Mobilités.

On constate que le nombre de lignes présentes dans OpenStreetMap a progressé lentement mais régulièrement pendant un an.
La communauté s'est notamment mobilisée à plusieurs reprises lors d'actions coordonnées pour accompagner les réorganisations de réseaux. On peut citer entre autres le réseau *Apolo 7* en avril 2018, des réorganisations diverses lors de la rentrée de septembre 2018 ou encore l'importante restructuration du réseau de bus parisien en avril 2019.

Le nombre de lignes dans OpenStreetMap a cependant atteint un palier depuis quelques mois : en effet, l'essentiel des lignes régulières sont maintenant présentes dans OpenStreetMap, seules les lignes peu fréquentes ainsi que les plus éloignées du centre de la région restent à cartographier.

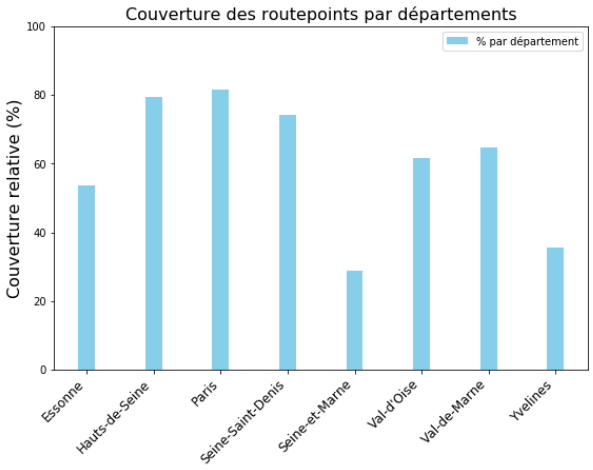
Nombre de routepoints

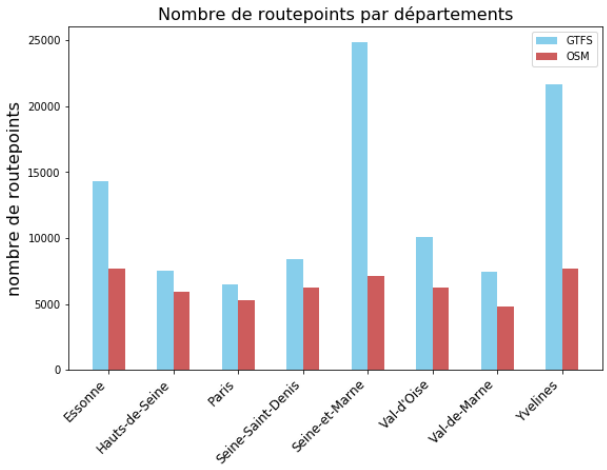
Pour mitiger les divergences de modélisation entre les deux sources, depuis avril 2018 notre étude s'articule largement autour du **routepoint**, c'est-à-dire un arrêt logique représentant un arrêt de bus desservi par une ligne dans une direction

donnée.

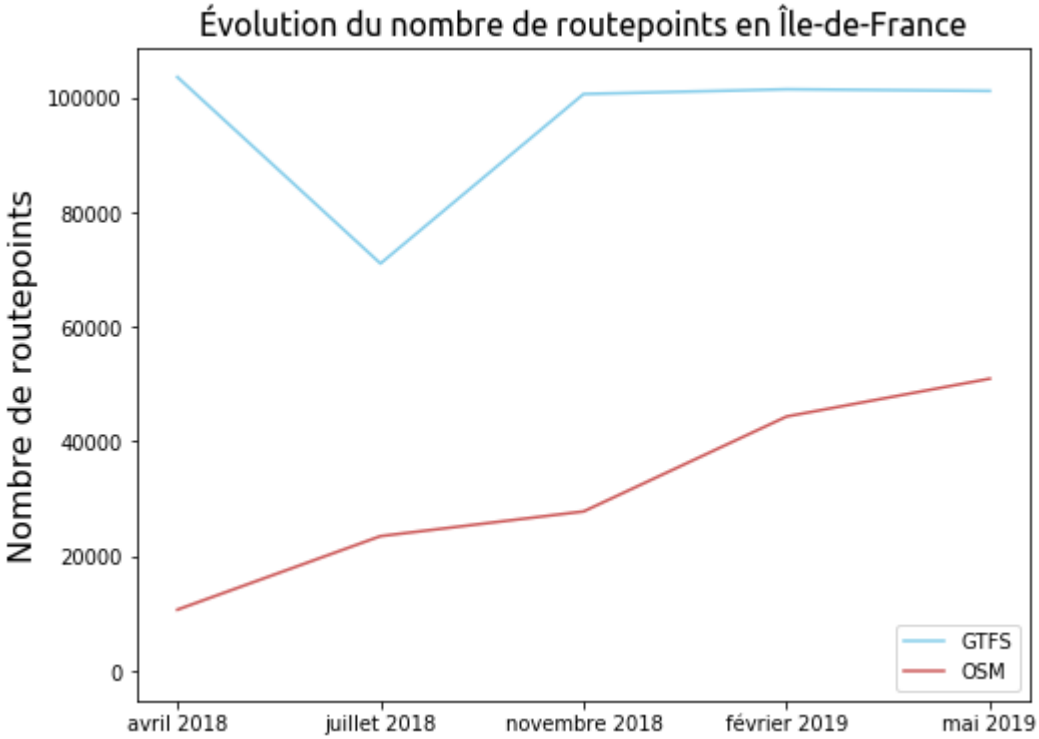
Voici la répartition du nombre relatif de routepoints d'OpenStreetMap par département.

Le graphique de gauche représente le pourcentage de couverture en routepoints de chaque département de la Région.
Le second graphique présente le nombre de routepoints de chaque département (en bleu dans l'open data d'Île-de-France Mobilités, et en rouge dans OpenStreetMap).





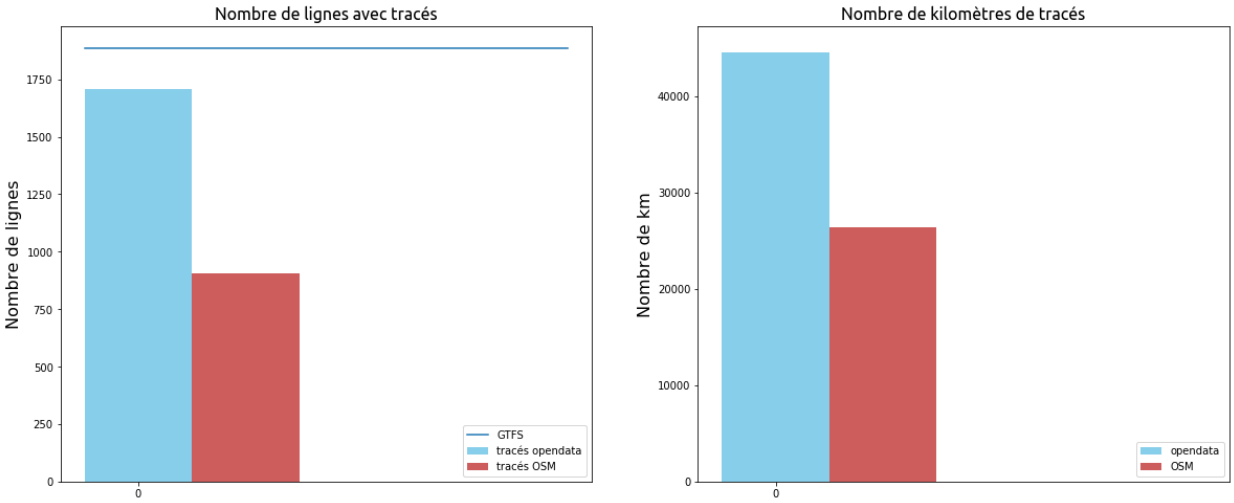
Le pourcentage global de couverture en routepoints sur la région Île-de-France est de **50 %**.
Voici un graphique de l'évolution du nombre de routepoints dans les deux jeux de données :



On constate que si le nombre de routepoints du GTFS est resté assez stable (à l'exception de la période des vacances scolaires d'été), le nombre de routepoints d'OpenStreetMap a en revanche été multiplié par 5 en un an, grâce à l'importante mobilisation des contributeurs.

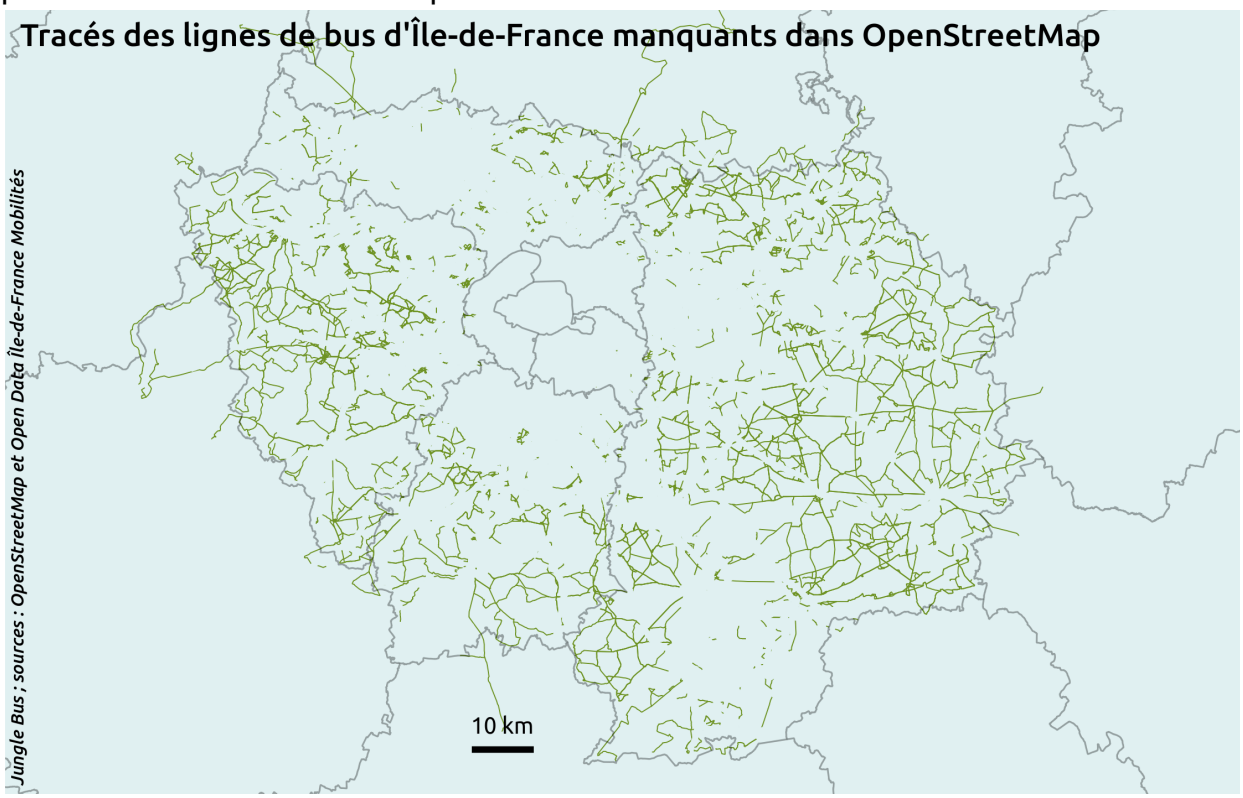
Nombre de tracés de lignes

Nous avons introduit, depuis l'audit de juillet dernier, un indicateur sur le nombre de lignes qui disposent d'un tracé, en nous basant exclusivement sur OpenStreetMap : en effet, c'était jusque là la seule base de données ouverte à disposer d'informations sur les tracés des lignes de bus.
Mais depuis le mois dernier, Île-de-France Mobilités met à disposition sur son portail open data un jeu de données décrivant les trajets géographiques des lignes (https://opendata.stif.info/explore/dataset/bus_lignes) régulières de bus de la région.
Environ 90% des lignes ont un tracé dans ce nouveau jeu de données. À l'inverse, OpenStreetMap est moins exhaustif, puisque seules 68% environ des lignes d'OpenStreetMap disposent d'un tracé.
Cela correspond respectivement à environ 45 000 kilomètres de tracés pour l'Open Data, contre environ 26 000 km pour OpenStreetMap.



Si on analyse géographiquement les deux sources, on constate que si les données officielles sont bien réparties sur toute la région, on ne peut pas en dire autant des données OpenStreetMap, dont la couverture est plus hétérogène.

Voici les zones où des tracés sont manquants dans OpenStreetMap mais bien présents dans les données Open Data officielles :



Comme lors de notre analyse de la répartition géographique des arrêts d'avril dernier, on constate que les quatre départements du centre de la région sont très bien couverts en données OpenStreetMap alors qu'à l'inverse il manque énormément de tracés de lignes dans l'Essonne, les Yvelines, le Val d'Oise et surtout la Seine-et-Marne.
Il existe une corrélation entre la couverture OpenStreetMap et la densité de

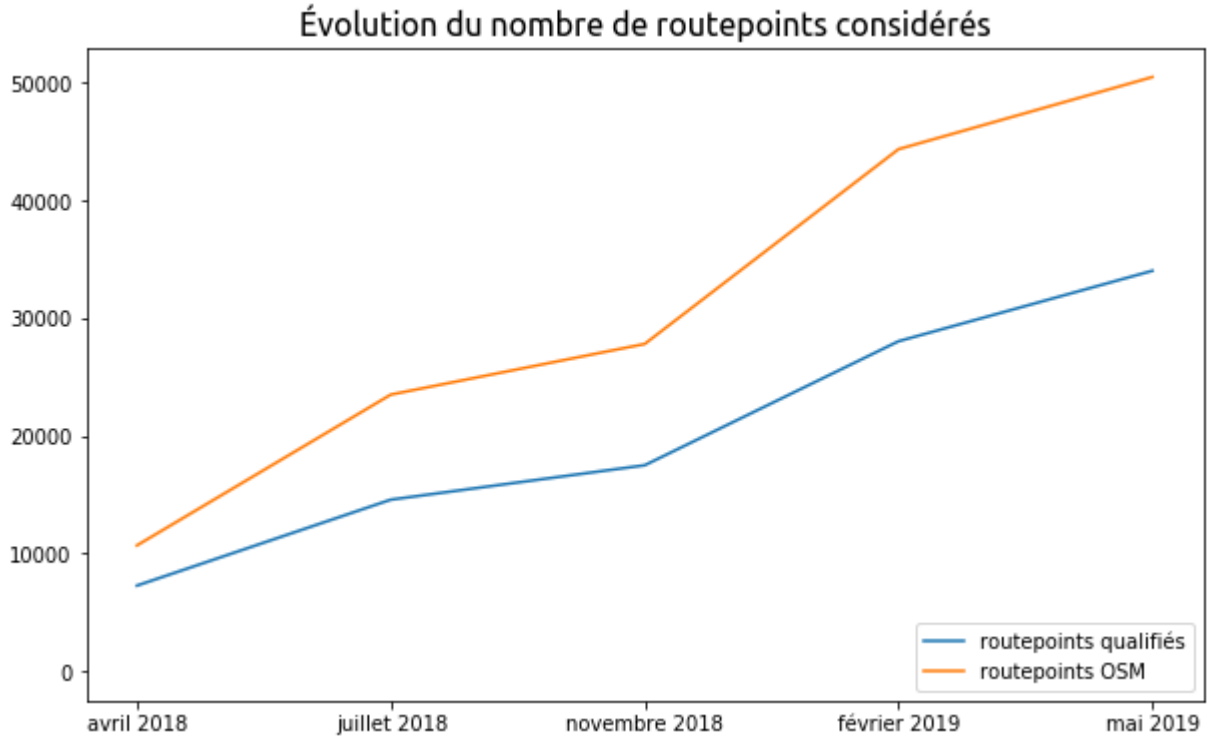
population.

Analyses qualitatives

L'objectif des analyses qualitatives est d'évaluer l'écart entre les deux sources sur un ensemble de critères, afin de déterminer dans quelle mesure il est possible d'enrichir les données Open Data officielles à l'aide des données OpenStreetMap.

Rappels méthodologiques : Nous ne conserverons que les routepoints (arrêt logique représentant un arrêt de bus desservi par une ligne dans une direction donnée) des deux sources que nous avons pu faire correspondre, en utilisant les référentiels d'Île-de-France Mobilités (REFLEX pour les arrêts, CODIFLIGNE pour les lignes) et une comparaison de chaînes de caractères pour les directions. Cette méthode est volontairement restrictive afin de limiter les faux positifs : seuls 34% des routepoints officiels sont ainsi comparés pour la suite de cette étude.

Voici un graphique de l'évolution du nombre de routepoints utilisés pour les analyses qualitatives :



On constate qu'au fur et à mesure de la progression de la cartographie dans OpenStreetMap, on a ainsi pu multiplier par plus de 4 en un an le nombre de routepoints analysés.

Les données brutes mises en correspondance et utilisées pour la suite de l'étude

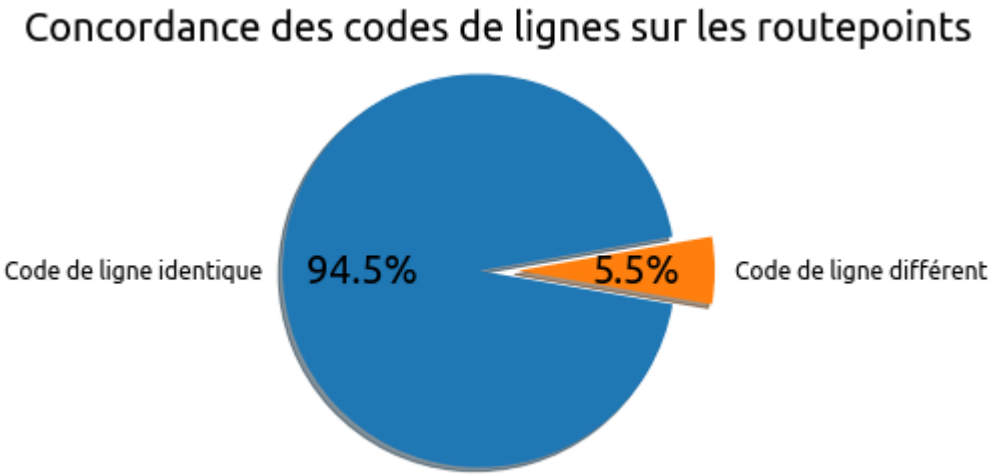
sont disponibles au téléchargement à cette adresse (https://raw.githubusercontent.com/Jungle-Bus/ref-fr-STIF/master/audit_routepoints/audits/2019_05/images/mapping_des_routepoints.csv).

Qualité des numéro de lignes

Lors de l'audit d'avril 2018, nous avons constaté

- qu'environ 95 % des lignes avaient un code de ligne identiques entre les deux sources
- qu'OpenStreetMap propose une dénomination plus proche de ce que constatera un voyageur sur le terrain dans les cas restants

Ces constats sont resté d'actualité depuis notre audit initial :

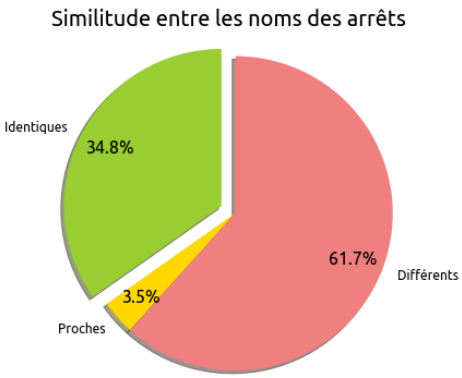


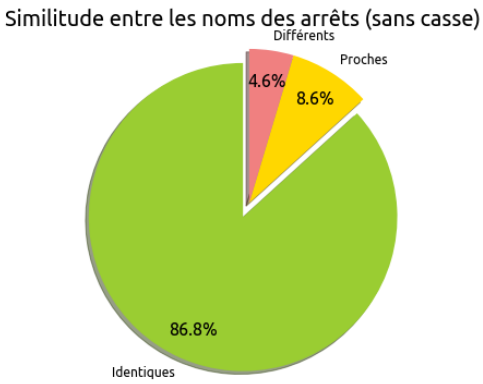
Qualité des noms

Lors de l'audit initial, nous avons constaté que les données du référentiel officiel sur le nommage des arrêts sont très hétérogènes et manquent d'uniformisation en comparaison avec celles d'OpenStreetMap.

Cependant, après un traitement de remise en cohérence de la casse sur les données Open Data, environ 95 % des arrêts avaient un nom identique ou proche dans les deux sources.

Sur ce sujet, nous constatons que ces constats restent d'actualité :



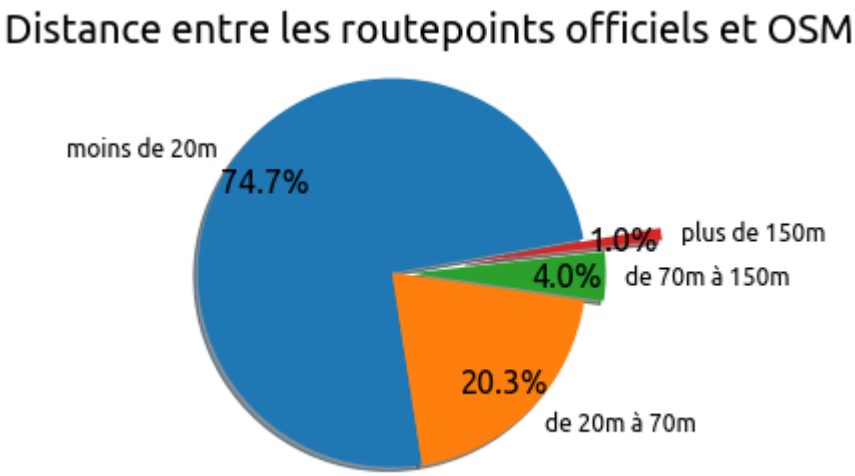


Qualité des positions

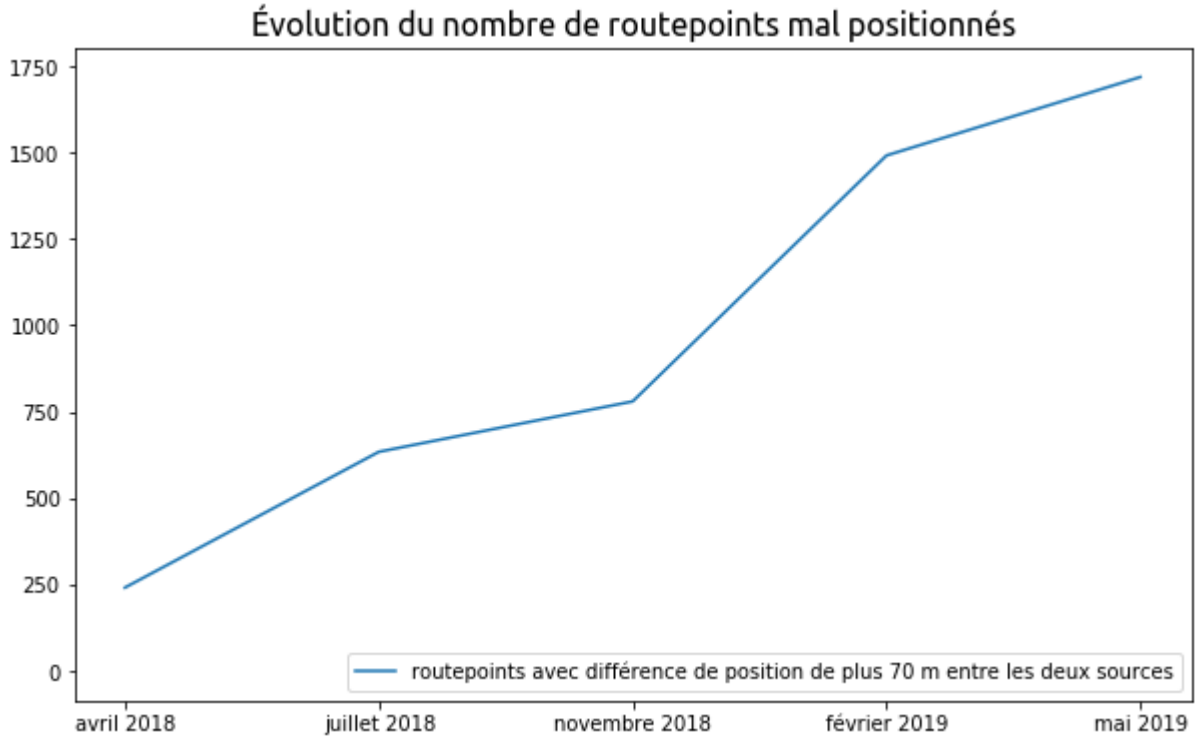
Lors de l'audit initial, nous avons constaté que l'essentiel des arrêts officiels sont positionnés à une distance raisonnable de leurs homologues d'OpenStreetMap.

Cependant, quelques centaines d'arrêts étaient situés dans le GTFS à plus de 70 mètres de leur version crowdsourcée par la communauté.

Ces constats se vérifient toujours :



Nous comptons à présent **5%** des routepoints pour lesquels les positions divergent de plus de 70 mètres entre les deux sources de données, ce qui correspond maintenant non plus à quelques centaines mais à plus de 1500 routepoints.

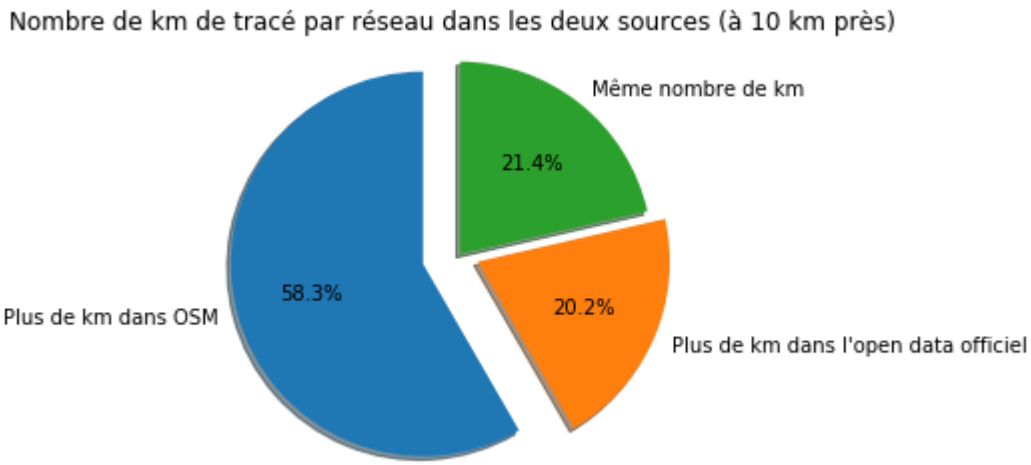


Enfin, afin de rendre possible l'amélioration des données officielles à partir d'OpenStreetMap, le détail des écarts de positions est disponible en téléchargement à cette adresse (https://raw.githubusercontent.com/Jungle-Bus/ref-fr-STIF/master/audit_routepoints/audits/2019_05/images/%C3%A9cart_positions_routepoints.csv).

Qualité des tracés de ligne

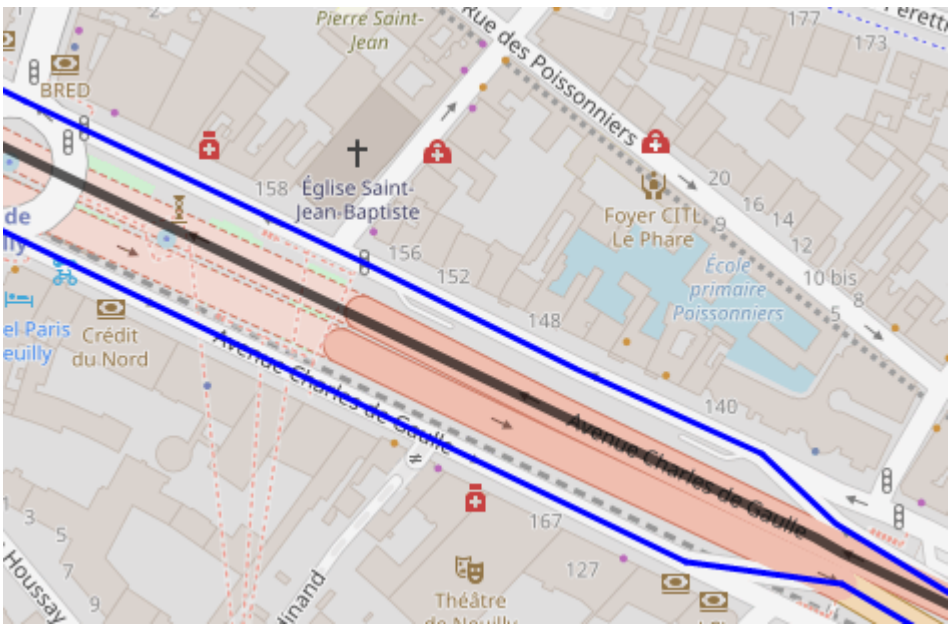
Dans la mesure où le nouveau jeu de données des tracés de lignes publié par Île-de-France Mobilités dispose d'un champ permettant de faire le lien avec le référentiel officiel des lignes, on peut faire correspondre les lignes des deux jeux de données afin de les comparer.
Après cette mise en correspondance, nous pouvons comparer 823 lignes (soit 43% des lignes officielles).

Comparons le nombre de kilomètres de tracés dans les deux sources pour chacun des 84 réseaux officiels représentés :



On constate que la plupart des réseaux représentés comptent plus de kilomètres de tracé dans OpenStreetMap que dans les données officielles, alors même que l'Open Data d'Île-de-France Mobilités totalise presque 2 fois plus de tracés.
Cela s'explique encore une fois par la divergence de modélisation entre les deux sources :

- les tracés que l'on trouve dans OpenStreetMap sont des tracés de parcours, c'est-à-dire d'un trajet effectué de manière récurrence par une ligne. On a collecté les tracés des différents parcours d'une ligne pour obtenir un tracé de ligne
- Il n'y a que peu d'informations concernant le procédé de création des tracés publiés par Île-de-France Mobilités, mais ils semblent être simplifiés et agrégés

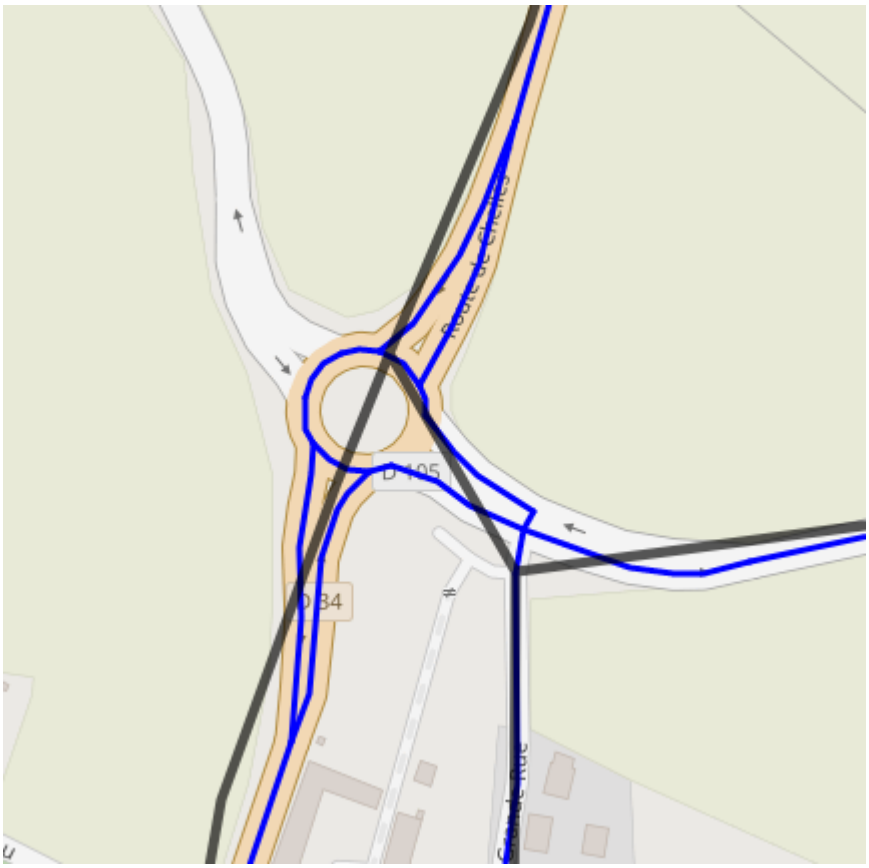


Ligne N24 du réseau Noctilien - données OSM et IDFM (fond de plan © les contributeurs d'OpenStreetMap) :
En noir le tracé officiel et en bleu le tracé d'OpenStreetMap : on remarque

qu'OpenStreetMap distingue le tracé du sens aller et celui du sens retour, tandis que le tracé d'Île-de-France Mobilité ne contient qu'un tracé de "tronc commun" unifié

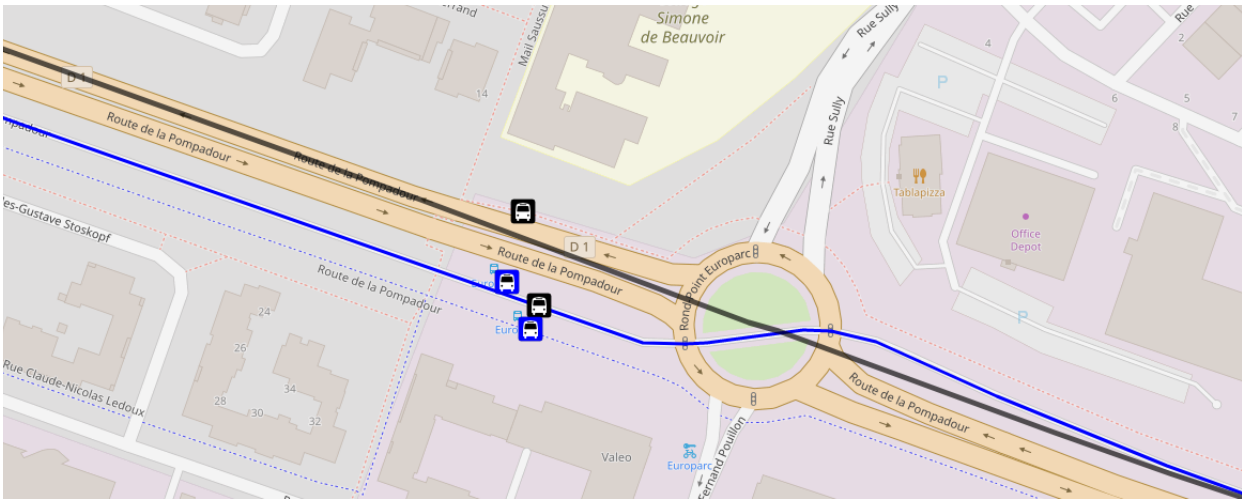
La différence de kilométrage peut également s'expliquer par la différence de précision :

- les tracés d'OpenStreetMap suivent parfaitement le filaire de voirie : en effet, un tracé dans OSM est constitué de la suite de voies empruntées. Par construction ce tracé est donc parfaitement conforme au tracé des rues
- les tracés d'Île-de-France Mobilités sont parfois simplifiés, ce qui réduit de fait leur longueur



Ligne 4s du réseau Apolo 7 - données OSM et IdfM (fond de plan © les contributeurs d'OpenStreetMap) :
En noir le tracé officiel et en bleu le tracé d'OpenStreetMap : on remarque que le tracé d'Île-de-France Mobilités ne suit pas très précisément les rues et est en conséquence bien plus court que celui d'OpenStreetMap

Enfin, comme précisé par Île-de-France Mobilités sur son portail Open Data, on constate parfois des tracés qui ne sont pas à jour avec les aménagements de voirie.



Ligne 393 du réseau RATP - données OSM et IdfM, fond de plan © les contributeurs d'OpenStreetMap :
en noir le tracé officiel et en bleu le tracé d'OpenStreetMap : le tracé (ainsi d'ailleurs que les positions des arrêts) d'Île-de-France Mobilités est très simplifié et ne tient pas compte de la voie de bus existante depuis 2011.

En conclusion, il est difficile de comparer objectivement les deux jeux de données car ils sont construits et modélisés de manière différente. Bien que les données officielles soient plus complètes, en terme de nombre de lignes mais aussi de détails sur la multitude des parcours et variantes de trajet, les données d'OpenStreetMap sont plus précises et fidèles à la réalité du terrain et parfois même plus à jour.

Focus sur deux réseaux

Afin de conclure notre série d'audits sur une note concrète, nous vous proposons de faire un focus sur deux réseaux d'Île-de-France en particulier et d'étudier plus particulièrement les similarités des deux sources.

RATP

Le réseau RATP est de loin le plus gros réseau de la région : environ 22% des routepoints officiels appartiennent à ce réseau.
Le nombre de routepoints d'OpenStreetMap est du même ordre de grandeur : le réseau est quasiment intégralement représenté dans OpenStreetMap.

La dénomination des réseaux y est en revanche légèrement différente : OpenStreetMap distingue par exemple les petits réseaux de proximité "Valouette"

ou encore “Autobus Suresnois”, qui sont opérés par la RATP, mais disposent d’une signalétique qui leur est propre. Dans les données officielles, ces lignes font toutes partie du réseau RATP.

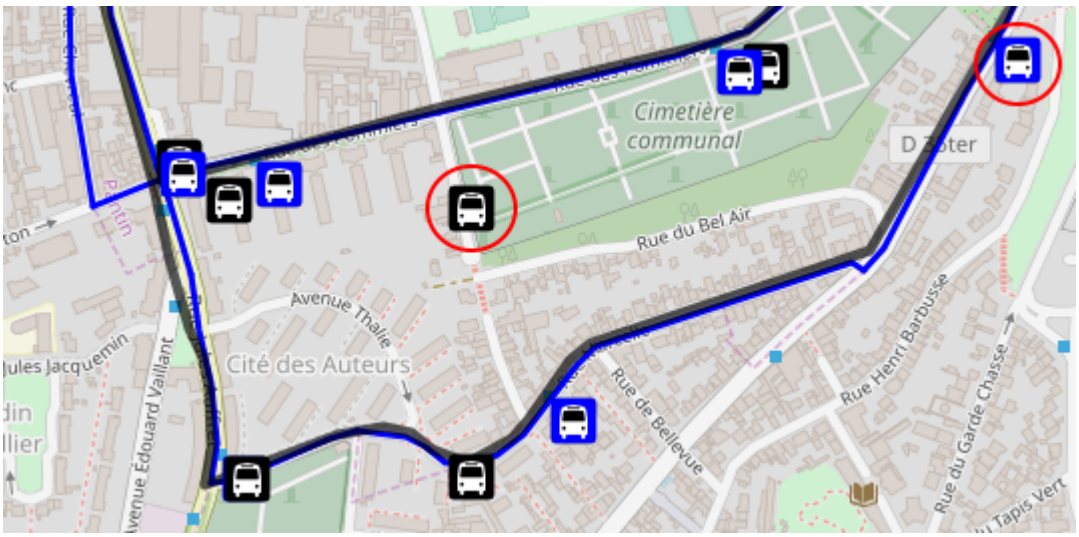
7 lignes gratuites desservent le Val de Bièvre. Elles disposent d’un logo aux couleurs de la communauté d’agglomération et sont rassemblées sous le nom “Valouette” . OpenStreetMap reprend cette dénomination dans le nom du réseau de ces lignes.

Comme pour le reste de la région, 95 % des codes de lignes sont identiques : les différences sont surtout des abréviations dans les données officielles (“CHOISYB” pour l’Open Data contre “Choisybus” pour OpenStreetMap).

Les noms des arrêts sont tous différents entre les deux sources : en effet, tous les arrêts RATP sont capitalisés et sans accentuation dans les données officielles. Cependant, après uniformisation de la capitalisation, on retrouve des noms très proches entre les deux sources, avec moins de 3% de divergences.

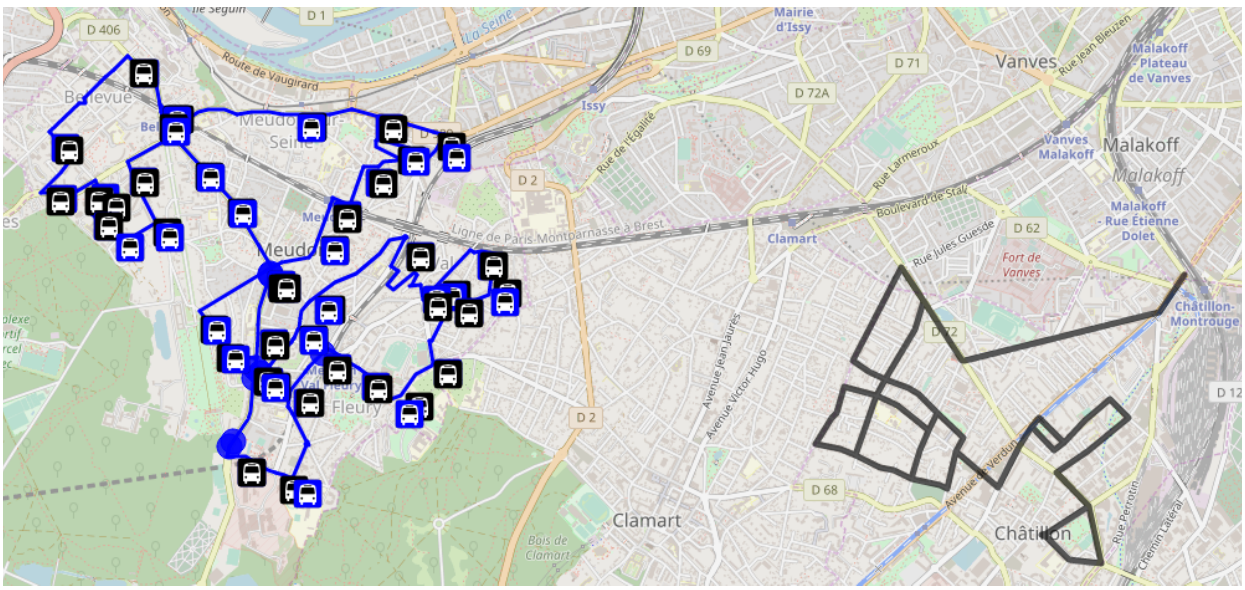
On constate que 95% des routepoints sont situés à moins de 70 mètres entre les

deux sources. La ligne 487 est celle qui comporte le plus de routepoints situés anormalement loin.



Ligne 330 du réseau RATP - données OSM et IdfM (fond de plan © les contributeurs d'OpenStreetMap) :
En noir le tracé et les arrêts officiels et en bleu le tracé et les arrêts d'OpenStreetMap : l'arrêt "Parc Henri Barbusse" officiel (au centre) est situé à plus de 400 mètres de son homologue crowdsourcé, à une position qui ne concorde pas avec les tracés de la ligne

Presque toute les lignes disposent d'un tracé dans OpenStreetMap et il existe encore une vingtaine de lignes pour lesquelles Île-de-France Mobilités n'a pas encore publié de tracés et où OpenStreetMap reste la seule source de données. C'est également le cas d'environ la moitié du petit réseau de transport de proximité Valouette.



Ligne TIM du réseau RATP - données OSM et IdfM (fond de plan © les contributeurs

d'OpenStreetMap) :
En noir le tracé et les arrêts officiels et en bleu le tracé et les arrêts d'OpenStreetMap : le tracé proposé par Île-de-France Mobilités ne correspond pas au trajet de la ligne.

Pays de l'Ourcq

Le réseau Pays de l'Ourcq, opéré par Transdev Marne et Morin, opère une vingtaine de lignes desservant principalement le nord-est de la Seine-et-Marne. Seules 3 de ces lignes sont existantes dans OpenStreetMap, ce qui constitue 57 routepoints (sur les 889 officiels).
Sur cet échantillon, nous ne constatons pas de divergence sur le nommage des réseaux et des lignes.

Concernant les arrêts, quelques petites différences apparaissent : quelques arrêts dans OpenStreetMap sont complétés par le nom de la ville afin de différencier des arrêts de même nom au sein d'une même ligne : "*Mairie (Varreddes)*" et "*Mairie (Congis)*" contre "*Mairie*" dans l'Open Data.

Enfin, les distances entre les arrêts dans les deux sources sont très proches avec au maximum une cinquantaine de mètres de différence.

Conclusion

Nous avons lancé cette étude en avril 2018 avec les objectifs suivants :

- évaluer la complétude des données OpenStreetMap
- mesurer les différences entre les deux sources de données sur un échantillon comparable
- évaluer le potentiel des données crowdsourcées d'OpenStreetMap pour enrichir les données d'Île-de-France Mobilités, en particulier sur la localisation des arrêts
- identifier les moyens d'actions pour augmenter la couverture OpenStreetMap dans le but de maximiser le potentiel d'enrichissement des données d'Île-de-France Mobilités.

Après avoir suivi l'évolution de nos deux sources de données, pendant un an, nous pouvons dresser les conclusions suivantes.

À ce jour, **OpenStreetMap ne contient pas l'intégralité des données transport**

disponibles sur la région : la contribution est géographiquement hétérogène et dès qu'on s'éloigne des tracés du réseau ferré d'Île-de-France dans les départements de la grande couronne, les données crowdsourcées sont incomplètes.
On note **une corrélation claire entre la densité de population et le taux de complétion des données dans OpenStreetMap**.

De plus, il existe de **réelles différences de modélisation** qui peuvent venir complexifier l'enrichissement d'une source à partir de l'autre : un arrêt, une ligne ou même un tracé de ligne ne représentent pas toujours le même objet dans nos deux sources. En s'appuyant sur des critères restrictifs, nous avons construit un échantillon comparable dans les deux sources afin de mitiger ce problème.

Sur cet échantillon comparable, les données des deux sources sont globalement identiques.
Même si les différences entre les deux jeux de données restent marginales, OpenStreetMap nous a permis d'identifier quelques cas où l'Open Data officiel pouvait être amélioré. En effet, **les données d'OpenStreetMap sont en général très proches de la réalité du terrain** et peuvent en conséquence proposer des corrections ou enrichissements sur des sujets utiles à l'information voyageur tels que le nommage des lignes et des arrêts en conformité avec la signalétique proposée ou encore la position des points d'arrêts.

Nous rappelons que les données d'OpenStreetMap sont en effet disponibles sous licence libre : il est possible d'extraire ces données à de multiples formats aussi bien pour des comparaisons (comme nous avons pu le faire ici) que pour enrichir et améliorer d'autres sources de données, y compris pour des applications de types cartographie ou calcul d'itinéraire.

Par ailleurs, en un an d'étude, le volume de notre échantillon comparable a été multiplié par plus de 4 ! Ainsi, si **la base OpenStreetMap n'est pas exhaustive, la contribution y est donc extrêmement active** : il y a donc un réel intérêt à s'appuyer sur la communauté pour améliorer les données de transports.

Enfin, notre étude focus sur 2 réseaux nous l'a montré, pour exploiter le plein potentiel des données crowdsourcées, il faut améliorer la complétude d'OpenStreetMap. Il serait donc pertinent de **terminer la cartographie de la région en concentrant les efforts sur les zones peu densément peuplées.**

Plusieurs axes nous semblent particulièrement pertinents à creuser pour dynamiser les efforts de la communauté :

- Soutenir le développement d'outils permettant de simplifier et de populariser plus largement les contributions sur les réseaux de bus. Le but est de faciliter la saisie des informations de lignes et de tracés mais aussi de suivre les évolutions d'un réseau. Des partenariats ponctuels avec les transporteurs ou les collectivités locales lors des re-structurations de réseaux ou des évolutions de voiries pourraient permettre de créer une donnée de qualité et plus en phase avec le terrain.
- Organiser des campagnes de contribution géographiquement ciblées en périphérie de région. Les zones restant à cartographier étant malheureusement plus pauvres en contributeurs, des partenariats entre la communauté et les transporteurs ou les collectivités locales pourraient s'avérer efficaces.
- Expérimenter de nouvelles approches aussi bien technologiques, en s'appuyant par exemple sur les dernières innovations d'analyse d'image et de détection automatisée pour prédire les positions des arrêts, que partenariales, en mettant en place une collaboration renforcée avec certains transporteurs en particulier.

Remerciements

L'équipe Jungle Bus remercie Cityway et Île-de-France Mobilités pour leur implication dans ce projet. Jungle Bus félicite également la communauté OpenStreetMap pour ses efforts continus de création et de maintenance des données de transport en Île-de-France.

Crédits

Cet audit a été réalisé par Jungle Bus (https://junglebus.io), grâce au soutien de Cityway (https://www.cityway.fr/), dans le cadre du projet m2i (http://www.mob2i.fr).

Contactez-nous à l'adresse contact-arobase-junglebus.io (http://contact-arobase-junglebus.io) ou via notre compte Twitter BusJungle (https://twitter.com/BusJungle).

(https://junglebus.io)

Les données utilisées pour cette étude sont :

- les données OpenStreetMap, sous licence ODbL (© les contributeurs d'OpenStreetMap), datées du 19 mai 2019
- les données GTFS (https://opendata.stif.info/explore/dataset/offre-horaires-tc-gtfs-idf/information/) mises à disposition en OdbL par Île-de-France Mobilités et datées du 17 mai 2019
- les focus ont été réalisés antérieurement au reste de l'étude, avec les données du début du mois de mai
- les tracés des lignes de bus d'Île-de-France (https://opendata.stif.info/explore/dataset/bus_lignes/information/) mis à disposition en OdbL par Île-de-France Mobilités en avril 2019
- le référentiel des lignes (https://opendata.stif.info/explore/dataset/referentiel-des-lignes-stif/information/) mis à disposition sous Licence ouverte par Île-de-France Mobilités afin de faire le lien entre les lignes du GTFS et celles du fichier des tracés, données datées du 19 mai 2019

Le code source utilisé pour préparer ces données et calculer les différents indicateurs présentés est consultable sur l'organisation Github de Jungle Bus (https://github.com/Jungle-Bus/ref-fr-STIF).