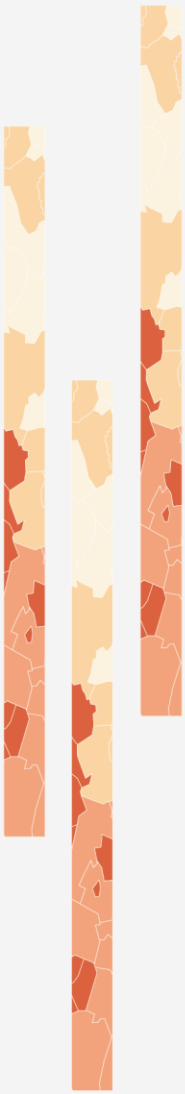
Généralités liées à OpenStreetMap et la complétude des données

Matthieu Viry, Timothée Giraud, Marianne Guérois, Ronan Ysebaert, Nicolas Lambert, Amel Feredj

RAPPORT TECHNIQUE

5. Généralités liées à OpenStreetMap et la complétude des données



RÉSUMÉ

Ce rapport technique présente, dans un premier temps, le projet OpenStreetMap (OSM) par son historique, sa philosophie et les applications qui en sont faites.

La deuxième partie aborde plus en détails la structure des données et les manières de les extraire selon le type de besoin.

Enfin la question de la complétude des données est abordée dans une troisième partie, présentant une partie de la riche bibliographie sur le sujet tout en abordant plus en détails les méthodes et résultats de plusieurs de ces études.

AUTEURS

Matthieu Viry, Timothée Giraud, Marianne Guérois, Ronan Ysebaert, Nicolas Lambert, Amel Feredj (UMS RIATE)

EXPERTS

Sophie de Ruffray (UMR IDEES), Grégory Hamez (LOTERR)

TYPOLOGIE SOCIO-ÉCONOMIQUE DES RÉGIONS FRONTALIÈRES DE L'UNION EUROPÉENNE (2000-2012)

Étude commanditée par le Commissariat Général à l'Égalité des Territoires

TABLE DES MATIÉRES

TABLE DES ILLUSTRATIONS

1 Le projet

1.1 La genèse

Le projet OpenStreetMap (OSM) est initié en 2004 par Steve Coast au Royaume-Uni (*University College de Londres*), notamment encouragé par le manque de données accessibles ou réutilisables. C'est un projet communautaire et sans but lucratif qui vise à permettre la diffusion de données géographiques sous licence libre et il s'inscrit à ce titre dans le courant du libre et dans celui du *web 2.0*.

La fondation OpenStreetMap est établit en avril 2006 pour encourager le développement et la diffusion de données géographiques (mais aussi gérer l'infrastructure matérielle du projet et le protéger juridiquement). C'est grâce à cette fondation que le projet peut recevoir des dons, qui sont ainsi répartis par la fondation selon les différentes nécessitées.

Dès décembre 2006 *Yahoo!* autorise l'utilisation de ses images aériennes en guise de fond de carte lors de saisies cartographiques sur OSM.

Par la suite, dès 2007, différents imports ont été réalisés par des structures publiques (bureau du recensement américain et sa *BD TIGER* par exemple) ou par des entreprises privées (notamment *AND* aux Pays-Bas).

L'association OpenStreetMap France a vu le jour en 2011, afin de promouvoir le projet et « *notamment la collecte, la diffusion et l'utilisation de données cartographiques sous licences libres* ». Cette association joue un rôle important dans la diffusion d'OpenStreetMap en France (par exemple via l'organisation d'événements divers : *carto-party*, édition française du *State of the Map*, journées *switch2osm*, et via la promotion et participation au projet de la Base Adresse Nationale en collaboration avec l'IGN, Etalab et la Poste).

En septembre 2012 un changement de licence concernant l'ensemble des données OpenStreetMap intervient et engendre une perte notable de données. En effet les données crées par des contributeurs dans le cadre de l'ancienne licence (*CC-BY-SA 2.0*) ont été supprimées lorsque leurs auteurs n'ont pas accepté les conditions d'utilisation de la nouvelle licence (***OdbL,*** *Open Data Commons Open Database License v.1.0 : http://opendatacommons.org/licenses/odbl/*). Cette perte de données est d'environ 1%, répartie de façon hétérogène sur le globe (certains pays tels que l'Australie, la Pologne et l'Égypte auraient ainsi été plus touchés).

On peut noter qu'en 2013 la société *Michelin Travel Partner*, une filiale de Michelin, a édité un plan de Clermont-Ferrand basé sur les données OSM et reprenant la symbologie graphique des cartes Michelin. La société a contribué à cette occasion à l'enrichissement de la base de données OSM[1].

En raison des conditions d'utilisation de l'API Google Maps (et notamment les modifications qu'elles ont connu en 2012) de nombreux sites internet se sont tournés vers OSM pour l'affichage de leurs cartes de localisation (*Craigslist*, *Flickr* et *FourSquare* par exemple).

En effet les données y sont ajoutées sous les termes de la licence ODbL mais on peut retrouver des données provenant d'autres sources compatibles (CC-BY pour l'Autriche, propriété de la DGFiP en France, *copyright* de la Couronne en Angleterre ou en Nouvelle-Zélande) et sous d'autres types de licence (licences propriétaires). Ces contributeurs spéciaux ne garantissent aucune responsabilité concernant la qualité de ces données. Des données ont également été fournies par la SRTM concernant la topographie terrestre, le *Corine Land Cover* et par différents types de partenaires concernant le réseau routier.

[1] Source: http://archive.wikiwix.com/cache/?url=http://www.sig-la-lettre.com/?OSM-met-la-gomme&title=[3]

La licence utilisée par OSM permet aux utilisateurs de récupérer, copier, utiliser transmettre et adapter les données en questions.

1.2 Les applications

Différentes applications des données issues d'OpenStreetMap existent : cartographie (« généraliste » ou thématique, avec mise en valeur de certains types de points d'intérêt comme les lieux accessibles aux fauteuils handicapés), routage/navigation (calcul d'itinéraire), etc. On note qu'OpenStreetMap n'empêche pas l'utilisation de leurs données à des fins commerciales (cf. http://wiki.openstreetmap.org/wiki/Commercial_OSM_Software_and_Services) et que des applications plus originales existent, telles que les simulateurs de vol *X-Plane* ou *FlightGear* qui utilisent notamment les données issues d'OSM pour la modélisation des différents éléments présents à la surface du globe.

Ces différentes applications possibles ainsi que l'originalité du projet (souvent présenté comme étant des projets les plus aboutit de VGI) ont donné lieu à de nombreuses publications scientifiques, particulièrement sur la thématique de l'évaluation de la qualité des données OSM par rapport à des données traditionnelles de référence (souvent propriétaires).

Parmi les originalités du projet OpenStreetMap on peut noter que la majorité des outils, et notamment des outils permettant de contrôler/vérifier les données, ne sont pas développés par la fondation OSM mais par des contributeurs/usagers (c'est par exemple le cas de *JOSM*, des exports réalisés par la société *GeoFabrik*, de l'*API Overpass*, etc.). L'abondance de projets n'étant pas gérés par la fondation est souvent présentée par les contributeurs comme étant une force du projet (et d'autre part la fondation n'a pas vocation à contrôler ce type de projets).

Un autre point à remarquer concerne la réactivité de la communauté OSM en cas de catastrophe humanitaire, dont certains membres s'attellent à cartographier rapidement des données à partir d'images satellitaires afin de fournir des cartes exploitables par les secours en place. C'est l'équipe HOT (*Humanitarian OpenStreetMap Team*) qui coordonne ces actions et qui fait le lien entre OSM et les humanitaires.

1.3 OSM en chiffres

Nombre d'utilisateurs	2 080 404 (1 million en janvier 2013)
Nombre de points GPS envoyés	4 620 511 896
Nombre de *nodes*	2 840 668 322
Nombre de *ways*	285 169 347
Nombre de *relations*	3 324 888

Tableau 1 - Données disponibles sur OpenStreetMap au 04 avril 2015

Le nombre d'utilisateur inscrit est important et semble en évolution constante (voir Tableau 1 et Figure 1).Ces données sont toutefois à relativiser, en effet en 2012 il est estimé (Neis et Zipf, 2012) que seulement 38 % des utilisateurs inscrits ont modifié ou créé un objet et que seulement 5 % des inscrits ont contribué « de manière significative ».

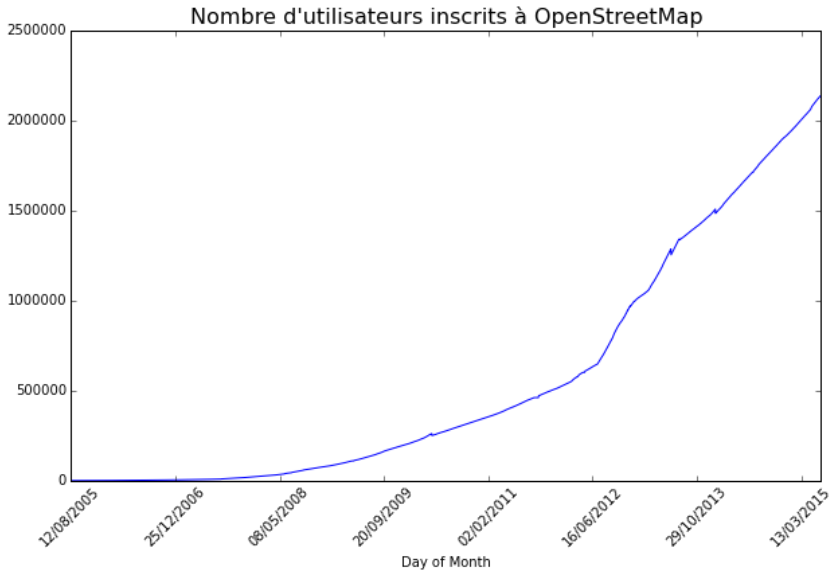


Figure 1 - Evolution du nombre d'utilisateurs inscrits au projet

2 Les données OpenStreetMap

2.1 Structure des données

OSM distingue 3 types d'***objets*** :

- ***node*** (un nœud) qui contient des données sur sa position (coordonnées au format WGS84), des métadonnées (la date de sa dernière modification et une référence au dernier contributeur l'ayant modifié ainsi qu'au lot de modification dans lequel il s'inscrit). Cet élément possède éventuellement un ou plusieurs tags sous la forme de pairs *key=value* (voir plus bas).

- ***way*** (une ligne) qui contient au minimum deux *nodes* et qui possède elle aussi des références concernant ses modifications ainsi qu'un ou plusieurs tags la décrivant. Cet élément représente notamment les rues, routes, rails, etc. Le sens de la ligne est décrit par l'ordre des points qui la composent. Toutefois, une ligne (*way*) décrivant un chemin fermé (c'est à dire que son premier et son dernier point sont identiques) et étant associée à des tags spécifiques (*multipolygon* ou *boundary* par exemple) décrit un objet surfacique.

- ***relation*** (une relation) qui décrit des relations plus complexe entre des objets : il peut s'agir d'une restriction spéciale lors d'un carrefour routier (*turn restriction*) comme d'une agrégation d'objets, hétéroclites ou non, pour délimiter un site précis (un aéroport pourra être décrit comme une relation, contenant l'ensemble des objets référencés lui appartenant : piste d'atterrissage, salle d'embarquement, etc.). Les relations contiennent donc des références à leurs membres, appartenant aux trois catégories citées (*node*, *way* et également *relation*) ainsi que leur rôle : au sein de la relation décrivant les limites communales de la ville de Strasbourg (*boundary=administrative*, *admin_level=8*), un membre est par exemple « *admin_centre* » et renvoi à un nœud correspondant à son hôtel de ville.

La base de données est récupérable dans son ensemble à partir d'un fichier généré chaque semaine structuré sous forme d'un fichier au format XML. Ce fichier (*planet.osm*) est très lourd (plus de 554GB en version XML non compressée, 39.6GB pour un fichier compressé *osm.bz2* et 26.7GB pour la version *.osm.pbf*)[2].

La structure des données et l'utilisation de *"tags"*

Chaque éléments est décrit par une ou plusieurs paires clé/valeur de la forme *key=value* (OSM propose 26 tags principaux, qui peuvent aussi être combiné à d'autres tags pour fournir des informations plus précises sur l'élément). Par exemple les principaux tags pour décrire un tronçon d'autoroute payant peuvent être :

highway = motorway
oneway = yes
maxspeed = 130
toll = yes
ref = A5

Il n'existe toutefois pas de restriction sur l'usage des tags et des valeurs (et certains éléments peuvent donc ne disposer d'aucun tag ou au contraire en avoir beaucoup). De plus, l'absence de spécifications prédéfinies lors de la saisie des tags OSM peut entraîner des fautes de frappes ainsi que des tags fantaisistes.

[2] Source : http://wiki.openstreetmap.org/wiki/Planet.osm (consulté le 18/05/2015).

Cette structure de données utilise le modèle de *Resource Description Framework* (RDF) décrit par Manela et Miller (2004). Ce modèle se base sur un ensemble de triplet (objet, sujet, prédicat) qui pourrait ici correspondre, dans le cas de notre route à (*point(x, y), highway*, *motorway*).

Les données suivants ce modèle RDF sont réputées difficiles à transformer en données orientées objet (Girres et Touya, 2010).

Enfin il faut noter que les données OSM ne sont pas directement adaptées à la mise en place d'un système de navigation. Il est nécessaire de sélectionner les lignes intéressantes/le réseau intéressant (via les tags appropriés) et d'en créer un graphe, prenant en compte la topologie de ces lignes (notamment pour préserver les sens interdits et le sens de circulation qui sont spécifié dans les tags OSM). Les données peuvent toutefois être qualifiées de « quasi-topologique » et c'est le stockage de relations topologiques explicites qui permet de bonnes performances lors de l'exécution de requêtes (Goodwin et al. 2008, in Girres et Touya, 2010).

2.2 La récupération des données brutes

Les données brutes (c'est à dire dans un des deux formats de référence OSM : *.osm* et *.osm.pbf*) sont exportées toutes les semaines vers un fichier *planet.osm* disponible sur différents miroirs (voir la liste sur http://wiki.openstreetmap.org/wiki/Planet.osm). Ce fichier est très volumineux et il existe plusieurs miroirs proposant des extractions réalisées à des emprises spatiales variées (continent, ville, etc.).

Des fichiers contenant seulement les différences (*changefile*) entre la version en ligne et les fichiers *planet.osm* sont générés quotidiennement (et chaque heure voir chaque minute depuis peu de temps) à l'échelle de la planète (et il est donc possible de trouver des serveurs proposant une version quotidienne du fichier *planet.osm* mis à jour par leurs soins). On note qu'un fichier quotidien de mise à jour pèse environ 40MB une fois compressé.

D'autre part des extraits régionaux sont fournis sur plusieurs sites : la société *Geofabrik*[3] propose des extraits à l'échelle des continents et des pays mis à jour quotidiennement et le site *MapZen*[4] propose quant à lui des extraits, mis à jour hebdomadairement, des principales métropoles du monde.

2.3 L'extraction / la préparation des données selon le type d'utilisation

Bien qu'elles soient basées sur un même fichier source (un fichier au format *.osm*, *.osm.pbf* ou *.osm.bz2*), les différentes utilisations possibles des données OpenStreetMap nécessitent des modes de préparation ou d'extraction parfois différents. En fonction de la complexité de la préparation, le volume des données obtenus peut-être très largement supérieur à celui du fichier d'origine (notamment lorsque qu'obtenus aux formats *osm.pbf* ou *osm.bz2* qui font tous deux appels à une forte compression). C'est par exemple le cas de la préparation nécessaire pour le routage avec OSRM (les données générées occupent environ 30Go, soit le double du fichier « Europe » d'origine ; elles comprennent notamment le graphe construit tout en conservant des informations sur les toponymes) ou pour la mise en place d'un outil de géocodage tel que Nominatim (sa mise en place pour l'Europe peut nécessiter plusieurs centaines de Go, notamment utilisés pour stocker de volumineux index permettant de garantir la performance de l'outil).

D'autre part, c'est à partir de ce même fichier que vont être réalisées les extractions thématiques, destinées par exemple à des fins cartographiques ou d'analyses spatiales. Ces extractions peuvent également s'avérer volumineuses en fonction de plusieurs facteurs tels que le nombre de champs attributaires créés et leurs contenus, le format de destination choisi et le niveau de détail (parfois élevé) des géométries à extraire. Ce type d'extraction vise généralement à passer de la

[3] http://download.geofabrik.de/
[4] https://mapzen.com/data/metro-extracts/

structure *key:value* d'OpenStreetMap à une structure tabulaire comportant les attributs choisis. A titre d'exemple, l'extraction du réseau routier seul[5] (sa géométrie et 3 attributs : son identifiant OSM, la valeur de clé *highway* et la valeur de la clé *maxspeed*[6]) occupe une table SpatiaLite de 3.25 Go après la création de l'index spatial.

Ces différentes utilisations nécessitent alors du matériel informatique adapté (forte capacité de mémoire vive et forte capacité de stockage, de préférence sur des supports rapides type SSD) afin d'obtenir des performances correctes lors des traitements ou bien simplement afin de rendre possible ces utilisations.

[5] Requête correspondant à *highway = motorway* | *trunk* | *primary* | *secondary* | *tertiary* | *unclassified* | *road* et ne prenant ainsi pas en compte les voies de services, chemins ainsi que certains types de rues des centres villes.
[6] Ce choix d'attributs arbitraire ne représente pas la richesse des informations présentes dans OSM et ne correspond également pas aux attributs retenus lors de cette étude, plus nombreux.

3 Analyses de la qualité des données

3.1 Les différentes formes prises par l'évaluation des données OSM

Une riche bibliographie scientifique est consacrée à OpenStreetMap, essentiellement dans le champ de la géomatique et de la géographie sociale. La majeure partie de ces études vise à évaluer la qualité des données disponibles via OpenStreetMap. Cette évaluation est généralement réalisée à partir d'un jeu de données de référence, pouvant être des données produites par une agence de cartographie d'État (BD Topo en France, BD TIGER aux USA, GIP en Autriche, etc.) ou de données commerciales produites par une entreprise privée (NavTeq, TeleAtlas/TomTom ou Google par exemple). Elle peut prendre différentes formes :

- évaluation de la **précision/justesse de la localisation des objets** : la localisation/géométrie des objets OSM est par exemple comparée à celle issue de la BD TeleAtlas (Zielstra et Zipf, 2010) : la localisation des objets TeleAtlas s'avère plus précise dans l'absolue, notamment en raison des techniques utilisées (combinaison de laser et GPS pour TeleAtlas permettant une précision inférieure à 1m, alors qu'OSM est essentiellement basé sur sur des relevé GPS traditionnels et sur de la localisation effectuées à partir d'image satellite géoréférencée), mais la superposition des deux jeux de données (avec un *buffer* de 10m) met en évidence plus de 80% de recouvrement pour les voies routières situées dans des grandes agglomérations. À l'inverse ce taux de recouvrement décroît (parfois de manière très importante) lorsque la comparaison est effectuée dans des petites/moyennes villes. Cette précision est également évaluée de manière plus fine (à l'échelle des nœuds du graphe routier), notamment pour évaluer les impacts d'une éventuelle fusion de jeux de données OSM/TeleAtlas (xxxx et al.). D'autre part, Girres et Touya (2010) relèvent une forte hétérogénéité dans le détail des objets géométriques, pouvant parfois conduire à une faible cohérence topologique entre certains objets ou limites administratives.

- évaluation **de la complétude « brute » des données** (comparaison de la masse d'information contenue) visant notamment à comparer la longueur de réseau référencé au sein d'une aire définie, en interprétant les différences entre les deux jeux de données comme étant des manquements au jeu de données dans lequel elles ne sont pas incluse. Cette comparaison brute (effectuée en 2010 sur l'ensemble de l'Allemagne par Zielstra et Zipf pour plusieurs dates entre 04/2009 et 04/2010) met en évidence, pour 2009, des manquements parfois important dans le jeu de données OSM. Toutefois ces manquements ont tendance à diminuer (pour l'Allemagne en prenant en compte l'ensemble des voies : 30 % de voies en plus dans *TeleAtlas* en 04/2009 contre moins de 5 % supplémentaire en 04/2010), notamment par l'ajout des rues situées au cœur des agglomérations et notamment les voies piétonnes, parfois peu renseignées dans les jeux de données propriétaires. La comparaison des différences concernant uniquement le réseau utilisable en voiture reste importante (40 % en Allemagne en 04/2010) mais le succès d'OSM, la poursuite des import de données de référence et la rapidité des évolutions de la base de données laissent présager une forte diminution de cet écart (et il est notamment relevé l'attitude des contributeurs qui, une fois la zone « quotidienne » cartographiée, vont poursuivre la prise de traces GPS dans des zones rurales, lors de déplacement en voiture ou lors de sorties et d'activités sportives, voir même en ciblant spécifiquement des zones qu'ils savent incomplètes). Girres et Touya (2010) ont effectué une comparaison des données OSM France avec la BD Topo au cours de laquelle il s'avère que de nombreux éléments présents dans la BD Topo ne sont pas présents dans OSM, les éléments concernés étant essentiellement des éléments de petite taille.

Cette comparaison OSM/jeu de données de références a été remise en application en juin 2011 en Allemagne, mettant ainsi en avant une différence de seulement 9 % entre le réseau *routable* extrait d'OSM et celui issu du jeu de données propriétaire. Lors de l'ajout et de la prise en compte de l'ensemble des rues, la comparaison penche alors en faveur d'OSM dont le jeu de données s'avère comporté 27 % d'informations supplémentaires.

- évaluation de la **complétude et de la justesse attributaire** (présence ou non d'attributs / si oui, attributs correctement utilisés). La justesse des attributs a également été comparée pour des points d'intérêts plus ponctuels que les routes tels que des commerces (notamment face aux données *NavTeq* et *Yelp* à Rome et à Londres, cf. Mashadi et al., 2012), révélant une très faible distance lexicographique concernant le nom de établissements ainsi que des différences notables concernant le type d'établissement rencontré (ces différences étant essentiellement dues au fonctionnement d'OSM qui n'impose pas de tag), les auteurs concluant ainsi l'article en qualifiant de « très positif le haut niveau de qualité dans OSM ».

Mashadi et al. (2012) testent également sur OSM les 4 principales propriétés applicables à Wikipédia en termes de qualité des articles[7], mettant ainsi en avant les points suivants :

- la nature des informations reportées dans OSM (beaucoup plus simple que la rédaction d'un article de qualité encyclopédique) permet parfois d'obtenir un haut niveau d'information dès la création de l'objet ; (→ ne valide pas la propriété énoncée pour Wikipédia) ;
- les utilisateurs qui fournissent de l'information de haute (ou de mauvaise) qualité tendent à continuer de fournir de l'information de haute (ou de mauvaise) qualité et d'autre part les usagers qui fournissent des informations complètes sur les objets ont tendance à fournir des informations correctes; (→ valide les 2 propriétés énoncées pour Wikipédia) ;
- il n'y a pas de corrélation entre le nombre d'édition effectué par un utilisateur et la qualité des objets qu'il modifie (→ ne valide pas la propriété énoncée pour Wikipédia).

D'autre part, différentes propositions ont pu être faite afin d'améliorer la justesse attributaire dans OSM, aussi bien par des méthodes « correctives » (Vandecasteele, 2013) telle que la modification d'attributs existants erronés par calculs de la distance sémantique par rapport aux attributs des autres objets, que par des méthodes préventives telles que la définition de spécifications claires/fermes ainsi que la mise en place de système de vérification à la volée de la cohérence des données saisies avec les spécifications en vigueur (Girres et Touya, 2010).

- **calcul des différences lors de calcul de chemins les plus courts**. Ce type de calcul/comparaison a notamment été effectué sur des calculs de trajets piétons (Zielstra et Hochmair, 2012), en milieu urbain, où il a pu être mis en avant la richesse du contenu OSM par rapport à des jeux de données traditionnels dont les informations sont destinées principalement au transport routier. C'est également ce type de méthode qui est utilisé par Graser et al. (2014) pour comparer la présence de *turn restrictions* entre 2 jeux de données.

- **« crédibilité » du jeu de données** OSM : la crédibilité est ici vu comme une nuance, volontairement plus subjective, aux mesures de la précision (*accuracy*) ou de la qualité d'un jeu de données; les auteurs notent que cette crédibilité « par la perception » peut être importante notamment pour ceux qui utilisent la BD OSM à des fins sociales/politiques/de communication (Flanagin et Metzger, 2008, p144). Les auteurs donnent les premiers éléments à prendre en compte pour évaluer la crédibilité d'un jeu de données issues de VGI, et des différents biais à prendre en considération (tendance à trouver Google Map plus crédible que OSM si on a jamais vu la représentation OSM, tendance des cartes à paraître plus crédibles (car on peut penser qu'elles représentent des éléments forcément objectifs) que d'autres formes de contenu généré par les utilisateurs, sensibilité aux erreurs/mauvaises informations différentes selon les usagers.

[7] C'est à dire : il existe une corrélation entre le nombre d'édition fait par un usagers et la qualité de ses contributions, l'ajout d'éléments structurant dans un article (lien, titre, etc.) indiquent un contributeur de qualité, les utilisateurs qui ont fait des contributions de haute (ou basse) qualité continuent à faire des contributions de haute (ou basse) qualité, un grand nombre d'éditions indique un article de qualité.

3.2 Les éléments récurrents identifiés

Plusieurs traits communs au jeu de données OSM ont été identifiés au cours de la réalisation des différentes études visant à évaluer sa complétude ou son adaptation à des usages particuliers :

- L'**hétérogénéité du jeu de données OSM** est presque toujours rappelée ou mise en avant dans les différentes études. En effet il n'existe pas de restrictions sur l'usage des tags et des valeurs (et certains éléments peuvent donc ne disposer d'aucun tag ou au contraire en avoir beaucoup). De plus, l'absence de spécifications prédéfinies lors de la saisie des tags OSM peut entraîner des fautes de frappes, voir des tags fantaisistes. A ces éléments s'ajoute l'absence de techniques de validation basé sur la qualité des contributions (c'est à dire par exemple qu'il n'y a pas de modérateur, et que les éditions supprimant volontairement des informations justes ne sont pas bloquées). Cette hétérogénéité peut s'expliquer par la coexistence de nombreuses sources de données et de nombreux modes de capture ainsi que par la diversité des profils des contributeurs. Ce phénomène est ainsi qualifié par Goodchild (2007) de « patchwork d'information géographique ».

- Plusieurs auteurs attestent d'**une plus forte qualité des données (justesse/complétude) dans les zones densément peuplées** (en terme de contributeurs OSM au moins) en comparaison à des zones peu peuplées et/ou situées en zone rurale (Neis et Zipf, 2010 ; Corcoran et Mooney, 2013), cette information étant notamment vérifiée en France où des départements comme la Creuse se sont avérés très peu complets[8] (Girres et Touya, 2010), confirmant ainsi la plus faible couverture des régions les moins peuplées (Haklay, 2010).

- On relève également que d'après plusieurs auteurs (notamment Zielstra, Hochmair et Neis (2013), Tenney (2014)) **la complétude et la précision des données augmente avec les imports en provenance de jeux de données de référence**. Cette information est toutefois à modérer en raison de l'attitude des contributeurs face aux données issues de base de référence (Zielstra, 2014 ; Neis, 2014), lesquels hésiteraient parfois à modifier des informations, même s'ils les jugent fausses, lorsqu'elles sont issues d'un jeu de données de référence. De même, bien que certaines données (présentes dans d'autres jeux de données), soient absentes d'OSM, il est parfois relevé que les modifications récentes des lieux (post-2008 par exemple), sont correctement/rapidement renseignées par les contributeurs (Hochmair et al. 2015, Transaction in GIS). Enfin, la question des imports externes est transversale avec les questions de licence : en effet des pertes d'information relativement importantes ont été observées lors du changement de licence d'OpenStreetMap[9] en 2012. Des manquements dans la cohérence ou dans la complétude du jeu de données ont pu être observés et directement rattachés à ces suppressions (Graser et al. 2014). Les modifications apportées par ces suppressions sont évaluables en étudiant l'historique des objets, conservé par OpenStreetMap et disponible via l'API OSM ou au sein des fichiers *planet.osm* historiques.

- **La majorité des études concernant la justesse de positionnement et la complétude attributaires d'objets OSM se concentre sur les routes** ou des éléments qui en sont dérivés (chemins cyclables, zones et chemins piétons, etc.), à l'inverse les études axées sur les schémas de contributions prennent en compte les différents éléments édités par les contributeurs dont sont étudiées les modifications.

---

[8] On peut facilement miser sur le fait que les résultats d'une telle étude, effectuée en 2015, donnerai des résultats différents, notamment à en juger par la comparaison entre la vue OSM de la Creuse en 2010 proposée en illustration par les auteurs et la vue disponible aujourd'hui au même niveau de zoom (témoignant à la fois des modifications apportées au moteur de rendu et de l'ajout d'informations dans OSM). Cf. Figure 2.

[9] Ce changement de licence a été voté en 2009 par les membres de la fondation OpenStreetMap. La nouvelle licence (ODbL 1.0) ayant été trouvée plus sure et plus claire que la licence précédente (CC BY-SA 2.0), pour les usagers comme pour les contributeurs. Ce changement de licence à pris effet le 12 septembre 2012 et les données publiée auparavant restent dans le cadre de l'ancienne licence.

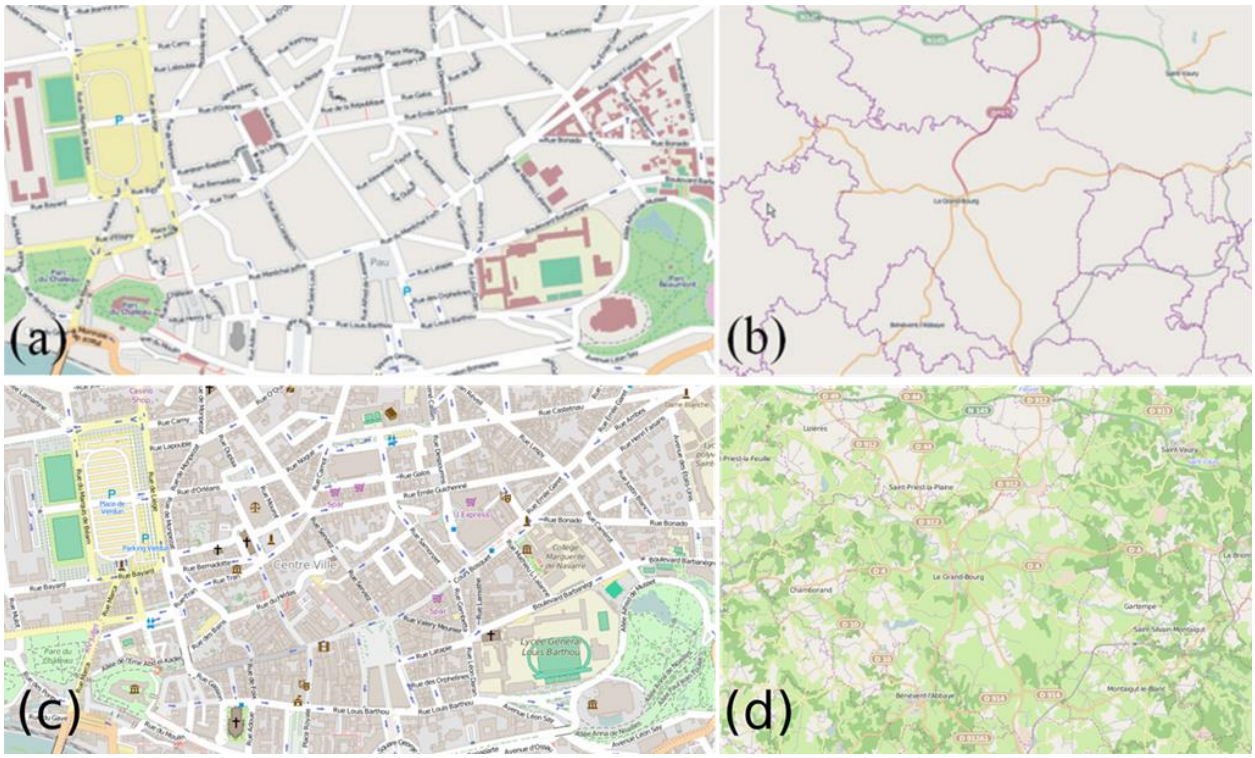


Figure 2 - Comparaison du niveau de détails d'OSM en France (2010-2015)

(a) Le centre de Pau en 2010 (b) La creuse en 2010
(Source : OpenStreetMap et contributeurs, dans : Girres et Touya, 2010)
(c) Le centre de Pau en 2015 (d) La creuse en 2015
(Source : OpenStreetMap et contributeurs, 2015)

On peut toutefois constater qu'en terme de complétude des données, et malgré les traits communs précédemment énoncés, des schémas différents se dégagent dans certaines zones géographiques, par exemple entre l'Allemagne et la Floride : si en Europe le niveau de détails semble augmenter dans les zones urbaines/ densément peuplées (notamment par l'ajout de nombreux chemins piétons ou cyclistes), il toutefois relevé par Zielstra et Zipf, A., 2010 (*Quantitative Studies on the Data Quality of OpenStreetMap in Germany)* et Zielstra et Hochmair, 2012 (*Using Free and Proprietary Data to Compare Shortest-Path Lengths for Effective Pedestrian Routing in Street Networks*) qu'aux USA, comparativement aux jeux de données propriétaires, le niveau de détails des réseaux routiers issus d'OSM est généralement plus important dans les zones rurales que dans les zones urbaines (notamment en raison des apports massifs issus de la *BD TIGER/line*[10] la base de données routière du bureau de recensement des États-Unis qui référence de nombreuses petites routes dans les zones rurales/agricoles).

3.3 Autres observations et focus sur quelques méthodes et résultats

Hawklay (2008) relève la **variabilité de la complétude et de la justesse de positionnement** des éléments issus du jeu de données OSM, ainsi qu'une **forte hétérogénéité en fonction des pays et en leur sein**. Toutefois il note également que, contrairement à d'autres jeux de données, les modifications sont ici simples à effectuer et rapidement prises en compte (il ne dépend ainsi qu'à celui qui remarque une erreur de la corriger).

De plus il existe des outils permettant aux non-initiés de laisser des notes aux contributeurs actifs afin de signaler des erreurs ou des modifications devant être apportées.

[10] TIGER: Topologically Integrated Geographic Encoding and Referencing.

Enfin Hawklay note que malgré ces imprécisions, le jeu de données n'en reste pas moins adapté à des nombreuses applications cartographiques/SIG (et il convient alors de définir quelles informations sont utiles selon les types de besoins, dans quelle mesure cette information est utile pour le besoin en question, et à partir de quel point il est possible de considérer que la couverture de l'information en question est suffisamment bonne).

Plusieurs auteurs associent l'évaluation de complétude et de la justesse des données avec les « **schémas de contributions** » suivis par les contributeurs. La compréhension de ces schémas (notamment via le nombre d'objets édités (dans l'absolu ou à chaque lot d'édition), la fréquence de contribution, le type d'édition (ajout, modification ou suppression) effectué, l'emprise des modifications de chaque lot, etc.) est importante afin de saisir la cohérence du jeu de données (et ainsi de comprendre certaines erreurs ou certains manquements). Les contributeurs peuvent ainsi être différenciés selon leur expérience (via le nombre de nœuds créés par exemple, cf. Neis et Zipf, 2012) ou selon le profil de leurs contributions (type d'édition, type d'objet (selon la clé ou selon la clé:valeur) le plus souvent édité, aire géographique de leurs contributions, etc., cf. Zielstra, Hochmair, Neis et Tonini 2014). L''étude de ces profils de contribution, basée notamment sur l'emprise géographique, se fait avec le postulat que des profils d'éditions différents vont apparaître suivant que les contributions soient réalisées à l'intérieur ou à l'extérieur de l'aire de contribution usuelle/résidentielle de l'utilisateur en question (*'home area'* dont l'emprise est définie au préalable et dont la validation des résultats a été soumise aux contributeurs concernés). Cette définition d'une aire de contribution "usuelle" est ici importante car elle permet de mettre en valeur l'effort de cartographie fournit suivant qu'il se situe à l'intérieur ou à l'extérieur de la zone, et elle permet également de supposer (conformément à d'autres études) que ces zones de contributions usuelles peuvent révéler des zones dans lequel le jeu de données s'améliore (correspondant alors ici aux zones urbaines et les plus densément peuplées, identifiées par d'autres auteurs comme étant les zones les plus riches en contributions). Dans cette étude il s'avère que les contributeurs éditent une variété d'objet plus large dans leurs zone de contribution usuelle (*home area*) et réalisent un effort de cartographie (en nombre de jour par exemple) plus important dans cette zone qu'à l'extérieur (*external area*). Ces constatations viennent conforter la dimension "connaissances locales" citée comme étant un avantage des projets de VGI.

L'étude de la complétude attributaire basée sur une approche prenant en compte les profils de contribution des utilisateurs par Mashhadi et al. (2012) met également en avant **le caractère consciencieux d'une majorité des contributeurs**, en faisant ainsi de bons fournisseurs d'information.

Le **département de la Sarthe** à fait l'objet d'une **évaluation de la qualité de son réseau routier** par des membres de l'École Supérieure des Géomètres Topographes (Petit et al., 2012). Cette étude, dont les observations ont été effectuées à deux reprises (janvier 2011 et janvier 2012), met en avant une hétérogénéité très importante au sein du département, qui respecte assez généralement le différentiel urbain/rural déjà observé. L'évaluation des différences est effectuée entre le réseau routier du RGE et celui d'OpenStreetMap. Ce dernier ne représente, en janvier 2011, que 15 % du RGE du département e la Sarthe. Cette exhaustivité tend toutefois à s'améliorer rapidement : en 01/2012 il s'agit de 35 % du RGE qui est représenté par OSM (masquant ici aussi une forte variation entre les différents lieux : 80 % de la ville du Mans est par exemple couverte). Les attributs observés sur le réseau routier sarthois sont globalement moins nombreux dans OSM que dans le RGE, toutefois les auteurs notent le nombre important de routes secondaires, tertiaires et résidentielles disposant d'informations (nom de la voie, référence, etc.), supérieure au RGE. La sémantique est également comparée, laissant apparaître que 71 % des noms de voies utilisés étaient identiques, les erreurs concernant alors souvent des différences orthographiques ou de ponctuation.

Corcoran et al. (2013) analysent la croissance du réseau de rues dans 3 villes différentes en Irlande. Les auteurs identifient deux phases successives dans l'accroissement de ce réseau, la densification puis l'exploration, qui désignent respectivement l'augmentation locale de la densité du réseau de rues et l'expansion du réseau vers de nouvelles zones.

Les auteurs explorent également l'évolution de la topologie du réseau (via l'utilisation de l'*index alpha*), et constatent une amélioration de cet indicateur (qui va de pair avec l'ajout de nouvelles arrêtes au graphe ainsi qu'avec leur validité topologique).

Les conclusions de l'article évoquent le rôle dominant de la densification sur l'exploration (phénomène qui va ainsi de pair avec le phénomène précédemment identifié et liant augmentation du volume de données OSM et zones densément peuplées). Enfin il est relevé que ce rapport entre densification et exploration n'est pas stable dans le temps, en effet le nombre de tronçons présent sur le réseau routier à un instant *t* est fini et la densification ne peut avoir lieu indéfiniment (tout comme l'exploration, dont la nature peut expliquer sa part moins importante, cette tache ayant déjà probablement été partiellement effectuée).

Pell et al. (2013) comparent les données OSM, le réseau routier du *Graph Integration Platform* (un projet soutenu par plusieurs agences d'État) et les données *TomTom MultiNet* en simulant la réalisation d'un problème de localisation optimale pour l'implantation de 10 plateformes de redistribution d'un fournisseur en agro-alimentaire (prenant ainsi notamment en compte la localisation de ses supermarchés clients, la possibilité de circuler en poids-lourds, etc.). Cette analyse a été menée en parallèle sur les 3 jeux de données et à fournit les mêmes résultats en terme d'implantation optimale des plateformes. Des tests réalisés à échelle plus fine, consistant à calculer des itinéraires précis via l'outil de TomTom, via la VAO (basé sur le GIP) et via OSRM ont pu montrer que les 3 avaient utilisé des parcours incorrects.

Différentes approches d'évaluation de la qualité des données OpenStreetMap sont conduites par la laboratoire de recherche en géomatique de la *Memorial University of Newfoundland (CA)*, notamment motivées par le fait que les méthodes utilisées dans des études précédentes sont parfois peu adaptées aux contenus générés par les utilisateurs.

Bégin et al. (2013) font remarquer que peu d'organisations utilisent les données issues de la VGI alors que de nombreuses d'entre elles en ont besoin. Cela s'expliquerait notamment par l'absence de méthodes fiables pour évaluer leur qualité (Van Ecel et al. 2010). Selon les auteurs, des éléments spécifiques aux VGI devraient être pris en compte dans l'évaluation de la qualité de ses données. Ces éléments, énoncés par Goodchild et Li (2012), reposent sur l'effet du nombre (*crowdsourcing approach*) et sur la confiance des contributeurs OSM (*social behaviour approach*).

Il est également relevé (Bégin et al. 2013) que **le choix des éléments édités par les contributeurs n'est pas due au hasard mais est souvent basé sur des préférences personnelles ainsi que sur la proximité spatiale**. Après avoir identifié des contributeurs et leurs éléments de contribution préférés, ils constatent la tendance qu'ont les sessions successives d'éditions des contributeurs à se superposer partiellement (pour venir par exemple ajouter des éléments d'importance moindre sur une zone déjà éditée) tout en s'étendant (pour venir ajouter des éléments parmi leurs éléments de contribution préférés). Les auteurs remarquent également que la cartographie ou la saisie d'éléments nécessitant une connaissance locale sont des événements plus difficiles à prévoir en se basant sur les préférences des contributeurs.

Graser A. et al. (2014) proposent de nouvelles méthodes de comparaison de réseaux routiers issus d'OpenStreetMap et d'un jeu de donnés de référence. Ces méthodes s'appuient sur des outils libres et largement diffusés dans le domaine des SIG, tel que la boite à outils d'algorithmes *Sextante toolbox* (qui a été enrichie de plusieurs outils, développés par les auteurs dans le cadre de la rédaction de l'article). Ces analyses visent à vérifier plusieurs aspects : la justesse de la localisation, la comparaison des longueurs de réseau, la complétude des attributs, la comparaison des restrictions de parcours (*turn restriction*) et la comparaison des informations concernant les voies à sens unique. Des informations sont également fournies sur la manière dont sont présentées les informations utile pour le calcul d'itinéraire dans les deux jeux de données.

- Les objets issus d'OSM s'avèrent être très correctement positionnés, en particulier pour les principaux couloirs d'autoroutes (p. 520), avec un défaut majeur de localisation concernant les bretelles d'accès.
- La longueur de réseau présente dans OSM s'avère comparable au jeu de données de référence (le niveaux hiérarchiques principaux (*motorway*, *trunk* et *primary*) possèdent des longueurs quasi-identiques, à l'exception des chemins piétons qui ne sont pas fournis dans le jeu de données de référence.
- La complétude des attributs s'avère largement inférieure dans OSM :
    - -78 % des voies de Vienne y possèdent un nom contre 94 % dans le jeu de données de référence,
    - -43 % des voies y comprennent des informations de vitesse contre 100 % dans le jeu de données de référence

- L'analyse des *turn restrictions* (basée sur des comparaisons de calculs d'itinéraire pour chaque *turn restriction* présente dans le jeu de données, et non pas sur la présence de cette information dans les attributs, faussée par les différentes entités pouvant être créées dans OSM) montre que 60% des *turn restrictions* présentes dans le jeu de données de référence figurent dans OSM.

- La comparaison des voies à sens uniques (elle aussi basée sur une méthode visant à réduire les différences liées à la manière dont sont fournies ces informations dans les deux jeux de données) montre que OpenStreetMap contient 87,9% des voies à sens uniques référencées dans le jeu de données de référence.

Ces différentes analyses ainsi que l'exactitude dans l'absolue des deux jeux de données a ensuite été vérifiée sur le terrain, dans des secteurs (0,5*0,5km) jugés représentatifs ou comportant beaucoup d'erreurs. Ces vérifications sur le terrain ont pu confirmer des absences dans le jeu de données OSM mais elles ont également montré, notamment dans un secteur comportant 21 erreurs, qu'une partie d'entre elles (8) venait du jeu de données de références alors que les autres (13) était issus de la non prise en compte d'un tag représentant une barrière au trafic routier (et donc une *turn restriction*) lors de la génération du graphe routier issu d'OSM.

La méthodologie mise en œuvre ici permet notamment de mettre en avant des zones comportant beaucoup de différences entre les sources de données (par exemple en vue de vérification sur le terrain). Cette méthodologie a aussi permis de mettre en avant que la majorité des erreurs concernent des voies de faibles importances. Ces erreurs risqueraient alors peu d'affecter le calcul d'itinéraires, celui-ci favorisant le passage par des grands axes de circulation.

Au-delà de fournir des informations sur le jeu de données OSM, cet article vise à proposer une méthodologie facilement reproductible (notamment car elle ne nécessite pas de compétences en développement ni l'utilisation d'outils propriétaires) afin des comparer les réseaux (routiers mais également piétons, cyclable ou ferroviaires) de différentes sources de données.

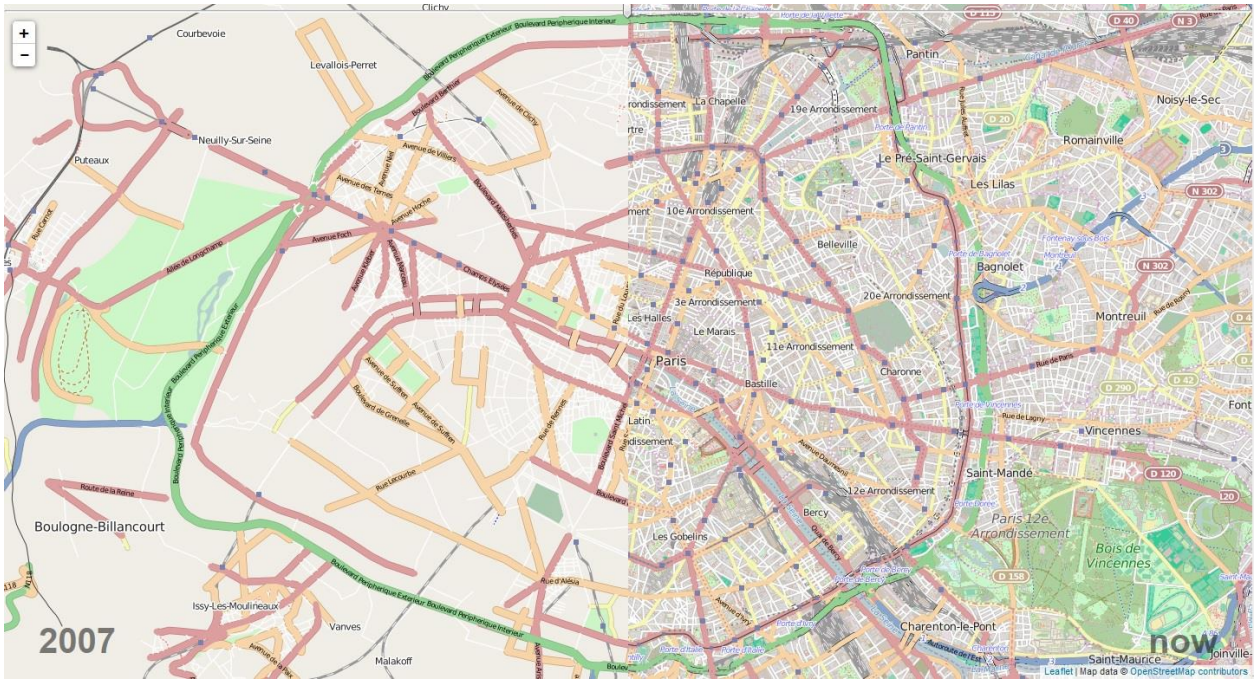


Figure 3 - Comparaison du rendu des données OSM pour la ville de Paris
(à gauche : juin 2007 – à droite : 17 août 2015)
Source : https://mvexel.github.io/thenandnow/#13/48.8618/2.3437

D'autres auteurs ont essayé d'évaluer la **complétude des données au OSM de manière conjointe à l'évaluation de la complétude d'autres jeux de données** (et non plus « par rapport à »). Pour ce faire, des données, considérées comme exhaustives, ont été récupérées (en l'occurrence le tracé des voies cyclables de 80 villes américaines, obtenu du croisement de plusieurs jeux de données et des données des municipalités concernées) et comparées à OSM et Google. Ces comparaisons, réalisées sur plusieurs villes et prenant en compte différents types de voies cyclables, montre qu'il n'existe pas un jeu de données parfait pour le calcul d'itinéraire à vélo (à titre d'exemple, au cours de cette étude, OSM s'avère plus complet que Google pour les chemins à Portland et pour les voies cyclables à Miami ; et à l'inverse Google dispose de meilleures informations pour les chemins/pistes à Miami).

Enfin plusieurs auteurs relèvent la **rapidité à laquelle grandit la base de données OSM** : le nombre de voies présent est en constante augmentation (des analyses temporelles locales montrent parfois des augmentations rapides, liées par exemple à l'organisation de *carto-party*) et le nombre de modifications subit par un objet tendrait à diminuer à mesure que sa description (*via* ses tags) s'approche de la réalité.

3.4 Observations sur la qualité des données réalisées au cours de l'étude transfrontalière

L'utilisation de différents types d'informations contenus dans la base de données OpenStreetMap a permis d'identifier différents éléments concernant d'une part la complétude du jeu de données et d'autre sa cohérence.

Le volume de données présent

Plusieurs points relevés par les auteurs précédemment cités peuvent facilement être vérifiés. Ils concernent notamment la différence, parfois marquée, entre les niveaux de détails présent dans les zones rurales et urbaines, correspondant souvent à la densité de population. L'importance de la densité de population sur l'ajout / la modification de données OpenStreetMap semble également être visible à l'échelle de l'espace d'étude, au sein duquel une comparaison, dans un maillage de

5km, des densités de population et des densités de routes met en avant le lien entre ces deux variables (Figure 5). A l'échelle des Etats, l'hétérogénéité relevée par certains auteurs est facilement perceptible avec une représentation en anamorphose (mettant en avant les pays dont le volume de données OSM est le plus important par rapport à leur superficie (en particulier l'Allemagne, la France et les Pays-Bas, voir Figure 4).

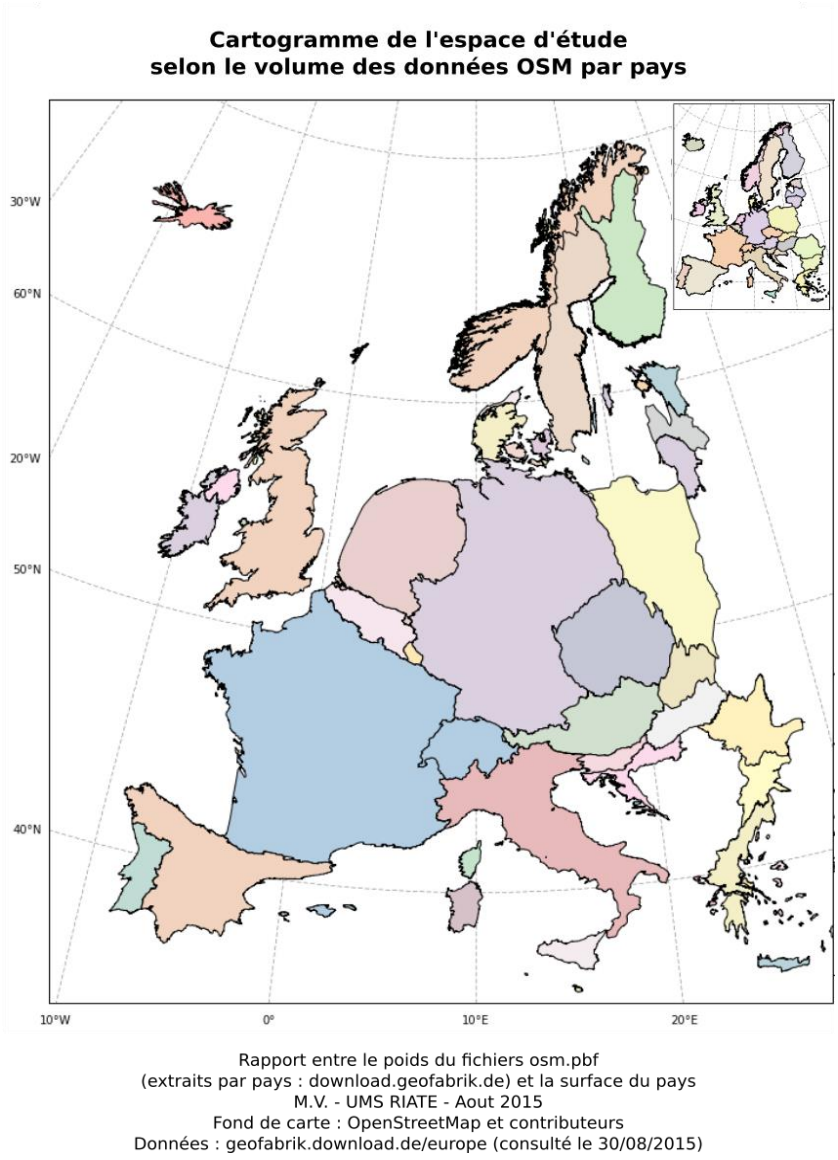


Figure 4 - Cartogramme de l'espace d'étude selon le volume de données OSM par pays

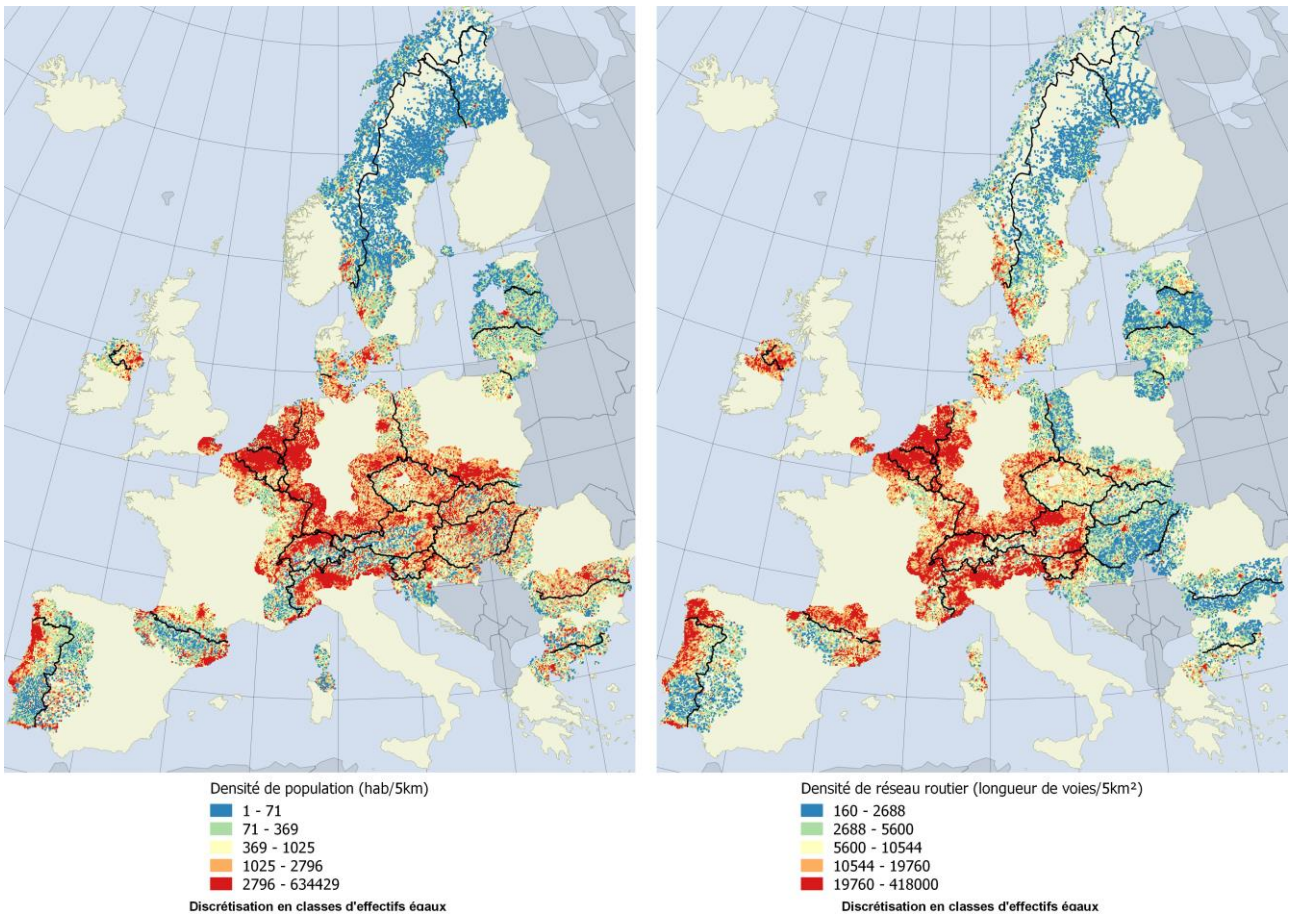


Figure 5 - Comparaison de la densité de population et de la densité de réseau routier (Maillage 5x5km – Données de population 2011 : grille GEOSTAT (Eurostat/EFGS) Réseau routier et tracé des pays : les contributeurs d'OpenStreetMap©)

Les résultats des calculs de temps de parcours

Les temps de parcours ont été calculés avec l'outil OSRM (présenté dans les fiches outils). Cet outils présente des particularités intéressantes (notamment en terme de performance) mais il n'est pas exempt de défauts. En effet il ne permet pas de prise en compte de la congestion et ne délivre ainsi qu'un temps de parcours théorique.

Des comparaisons ont été réalisées entre les temps de parcours obtenus avec OSRM et ceux provenant d'autres sources. Un premier échantillon de parcours réels issus du projet *EnviroCar*[11] a été récupéré et a été utilisé afin de mener une comparaison entre leurs durées / distances et ceux issus d'une part de l'API Google Direction et d'OSRM d'autres part ; le volume de parcours récupéré ne permettant pas de tirer de conclusion sur la validité des durées obtenues.

Deux échantillons issus de la base *METRIC*© (Mesure des Trajets Inter-Communes / Carreaux, mis au point par l'INSEE), et concernant des parcours au départ des départements de l'Ain (01) et des Hauts-de-Seine (92) ont également été obtenus dans le cadre d'échange avec le CGET. Les parcours au départ de l'Ain ont été utilisé afin de vérifier la validité des temps de parcours calculés avec OSRM, la comparaison ayant été menée en parallèle avec des temps de parcours obtenus via l'API Google Direction.

La comparaison a été effectuée sur les temps de parcours en minutes (et non sur les kilomètres de route parcourus) et en utilisant OSRM comme valeur de référence. Dans le cas de la base *METRIC*, c'est la valeur « *heure creuse* » qui a été utilisée.

[11] https://envirocar.org/

Les régressions linéaires des temps de parcours issus des couples de jeux de données OSRM / METRIC et OSRM / Google mettent en avant une forte cohérence entre les jeux de données concernés (voir Figure 6).

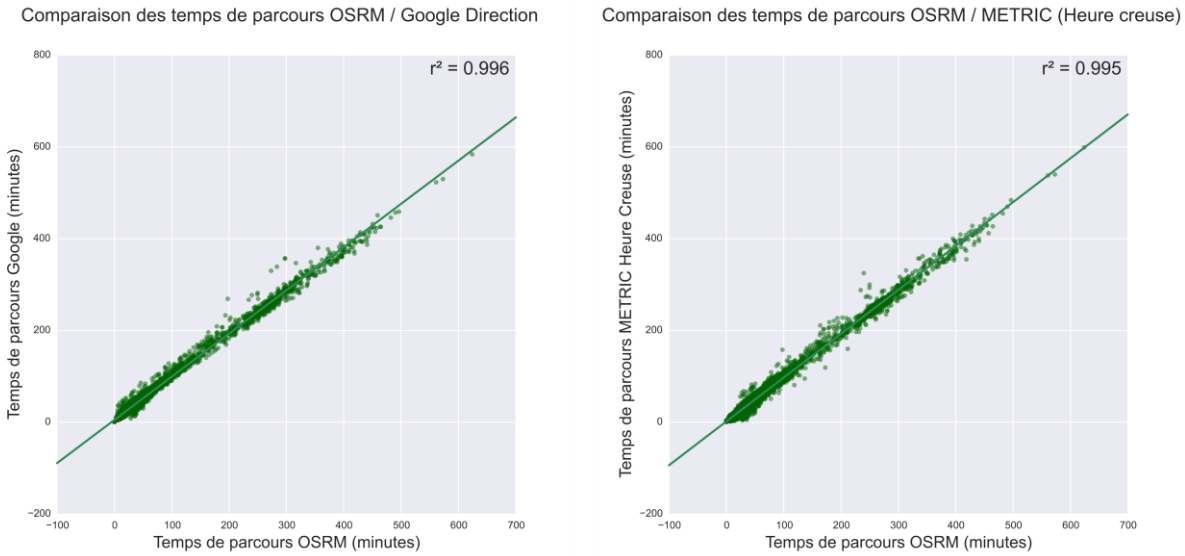


Figure 6 - Régressions linéaires issues de la comparaison des temps de parcours des 3 jeux de données

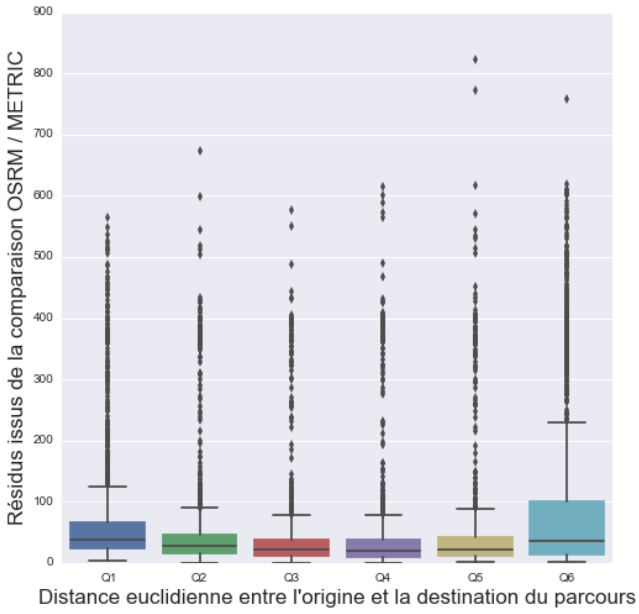


Figure 7 - Valeurs des résidus de la régression linéaire mis en relation avec la distance euclidienne entre le point de départ et le point d'arrivée.

L'observation des résidus issus de la régression linéaire entre les temps de parcours OSRM et METRIC met en avant qu'il ne semble pas exister de relation entre les forts écarts observés et la distance totale du trajet demandé. En effet, en discrétisant les valeurs prises par les résidus de la régression linéaire OSRM / METRIC, on constate que des trajets de longues distance ont pu être aussi bien surestimés que sous-estimés par OSRM (Figure 7).

Le même type d'exploration a été mené afin de tester l'existence d'une relation entre la valeur des résidus et la taille des communes d'arrivée ; l'échantillon de parcours considéré n'a toutefois pas révélé de relation entre ces deux éléments.

Parmi les données considérées, il est possible, afin de simplifier la lecture et l'interprétation des informations, de s'intéresser seulement aux parcours au départ d'une commune, notamment afin de visualiser les parcours présentant les plus forts écarts et afin de chercher des similitudes parmi eux (Figure 8).

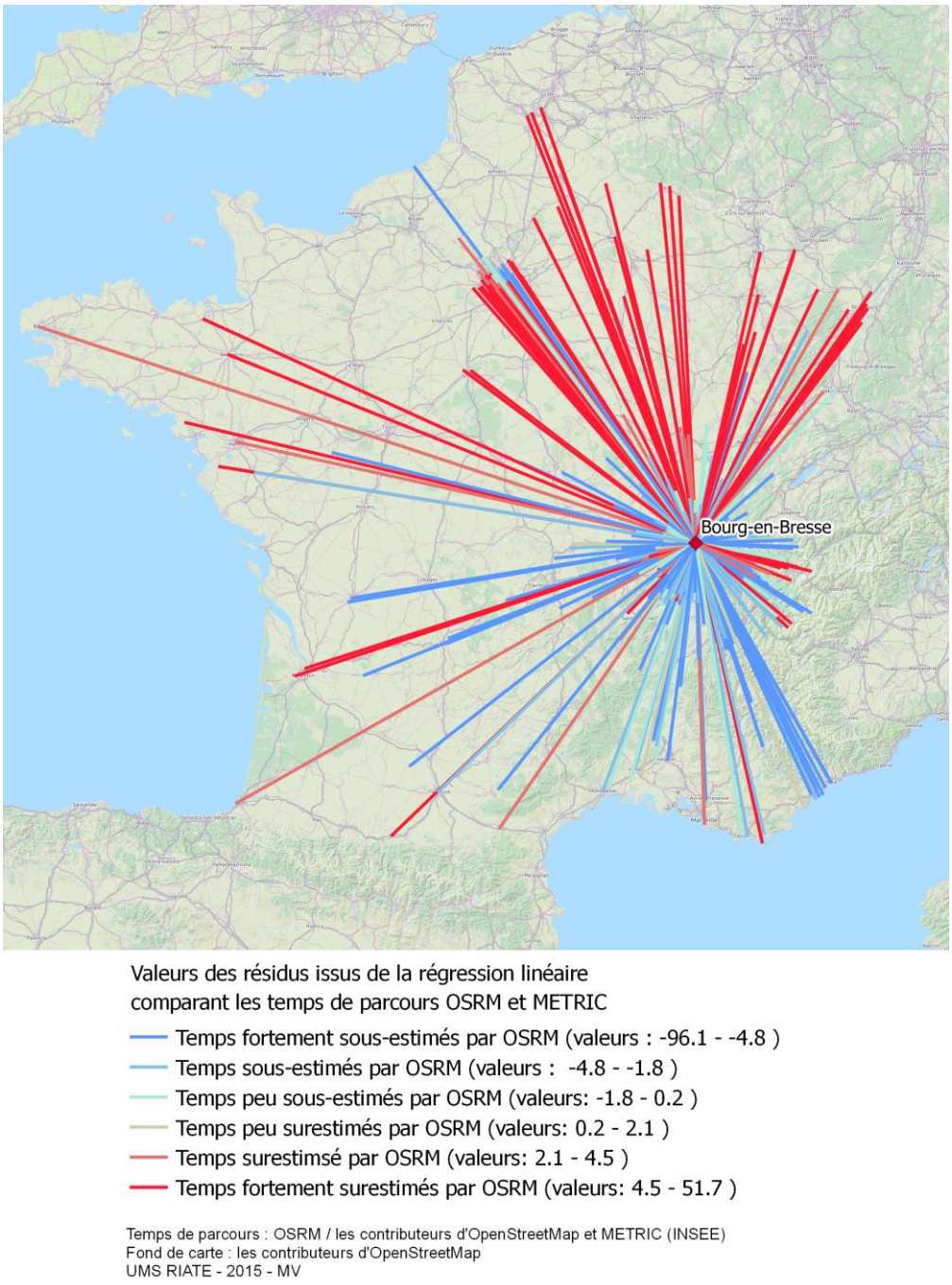


Figure 8 - Représentation des valeurs des résidus de la comparaison OSRM / METRIC au départ de Bourg-en-Bresse.

Les différences, bien que parfois marquées, entre les temps de parcours obtenus ne concernent qu'une minorité de parcours et ne semblent pas concerner un type de parcours facilement identifiable. Ainsi ces différences n'ont que peu d'impact sur le calcul des indicateurs présentés

dans les autres rapports techniques, notamment en raison du nombre important de parcours considérés pour le calcul de chacune des valeurs.

La présence de ces différences invite également à réfléchir aux types d'utilisation de chacune des bases de données. En effet, les codes-communes présents dans la base METRIC ont servi de référence aux requêtes effectuées auprès de Google et d'OSRM. Toutefois les coordonnées de ces communes (comme vu au dans le TR « Géocodage ») ne sont pas positionnées au centre des communes en question (alors qu'il est probable que le calcul d'itinéraire de METRIC utilise une position centrale dans les communes). Cette particularité a générée des erreurs qu'il peut être facile d'identifier (par exemple lorsque la localisation du point de départ ou d'arrivée nécessite un détour important afin de pouvoir atteindre une entrée d'autoroute, normalement accessible bien plus facilement depuis le centre de la commune considérée).

Ainsi les parcours issus de Google et d'OSRM sont précis (et prennent en compte les éléments du réseau présents sur les premiers/derniers kilomètres : sens-unique, bretelle d'autoroute, etc.) mais nécessitent de fournir des coordonnées (départ/arrivée), desquelles dépendent la qualité de trajet retourné. Au contraire, l'information obtenue via la distancier masque ces éléments, permettant une approche plus rapide lorsque l'échelle de travail est la commune française.

Chacun de ces outils présente des avantages ainsi que des limitations lorsqu'il s'agit d'obtenir des temps de parcours, ces éléments sont synthétisé dans le Tableau 2.

	OSRM	**METRICS**	**API Google Direction**
Prise en compte du trafic	Non (mais possibilité de préparer les données avec des données de congestion si disponibles)	Oui (sous forme « *Heure Pleine* » / « *Heure Creuse* »)	Oui (ajustée selon l'itinéraire et l'heure du parcours)
Disponibilité de la donnée	Données gratuites (réseau routier OSM) et libres/réutilisables.	Données payantes.	2500 requêtes gratuites par jour / Fortes restrictions sur la réutilisation des données.
Restitution du trajet emprunté	Oui (+ possibilité d'obtenir des trajets alternatifs / d'ajouter des points de passage intermédiaires)	Non, seulement distance et durée	Oui (+ possibilité d'obtenir des trajets alternatifs / d'ajouter des points de passage intermédiaires)
Matrice de temps de parcours	Oui	Oui, via jointures des tables voulues	Oui, limitée à 2500 requêtes
Utilisable à différentes échelles	Oui (continental à infra-communal)	Oui (seulement de communes à communes en France + infra-communal)	Oui (continental à infra-communal)
Multimodal	Oui (Voiture, vélo, marche)	Non	Oui (Voiture, vélo, marche, transports en communs et avions)

Tableau 2 - Synthèse des points forts et points faibles des outils comparés pour l'obtention de temps de parcours.

On note enfin qu'aucune de ces 3 bases de données ne permet, en l'état, de mener d'analyses prenant en compte la dimension temporelle. En effet, les données obtenues par la base METRIC sont figées à la date de leur création alors que les résultats provenant de Google et d'OSRM dépendent des informations présentes dans la base au moment de la requête.

Dans le cas de l'API Google Direction il ne semble pas exister d'option permettant d'obtenir l'itinéraire qui aurait été retourné il y a, par exemple, 5 ou 10 ans. Dans le cas d'OpenStreetMap la validité dépend de la date à laquelle la base de données a été récupérée. En l'état actuel de la complétude des données, et particulièrement lorsque le zone d'étude est étendue, il parait difficile de distinguer, dans le cas de l'utilisation d'une ancienne version de la base, ce qui relèverait de contributions (alors que les aménagements sont déjà existants) ou ce qui relèverait de réels ajouts/modification sur le réseau.

Bibliographie

Ali, A.L., Schmid, F., Salman, R. Al-, and Kauppinen, T. (2014). Ambiguity and plausibility: managing classification quality in volunteered geographic information. (ACM Press), pp. 143–152.

Bégin, D., Devillers, R., and Roche, S. (2013). Assessing volunteered geographic information (VGI) quality based on contributors mapping behaviours. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences *XL-2/W1*, 149–154.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C., and Foody, G. (2013). Using control data to determine the reliability of volunteered geographic information about land cover. International Journal of Applied Earth Observation and Geoinformation *23*, 37–48.

Corcoran, P., Mooney, P., and Bertolotto, M. (2013). Analysing the growth of OpenStreetMap networks. Spatial Statistics *3*, 21–32.

van Exel, M., Dias, E., and Fruijtier, S. (2010). The impact of crowdsourcing on spatial data quality indicators. In Proceedings of the Sixth International Conference on Geographic Information Science (GIScience 2010), (Zurich), p. 15.

Girres, J.-F., and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset: Quality Assessment of the French OpenStreetMap Dataset. Transactions in GIS *14*, 435–459.

Goodchild, M.F. (2007). Citizens as sensors: the world of volunteered geography. GeoJournal *69*, 211–221.

Goodchild, M.F., and Li, L. (2012). Assuring the quality of volunteered geographic information. Spatial Statistics *1*, 110–120.

Goodchild, M.F., Fu, P., and Rich, P. (2007). Sharing Geographic Information: An Assessment of the Geospatial One-Stop. Annals of the Association of American Geographers *97*, 250–266.

Graser, A., Straub, M., and Dragaschnig, M. (2014). Towards an Open Source Analysis Toolbox for Street Network Comparison: Indicators, Tools and Results of a Comparison of OSM and the Official Austrian Reference Graph: Towards an Open Source Analysis Toolbox for Street Network Comparison. Transactions in GIS *18*, 510–526.

Haklay, M. (2010). How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. Environment and Planning B: Planning and Design *37*, 682–703.

Haklay, M., and Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. IEEE Pervasive Computing *7*, 12–18.

Heipke, C. (2010). Crowdsourcing geospatial data. ISPRS Journal of Photogrammetry and Remote Sensing *65*, 550–557.

Hochmair, H.H., Zielstra, D., and Neis, P. (2015). Assessing the Completeness of Bicycle Trail and Lane Features in OpenStreetMap for the United States: Completeness of Bicycle Features in

OpenStreetMap. Transactions in GIS *19*, 63–81.

Mashhadi, A., Quattrone, G., Capra, L., and Mooney, P. (2012). On the accuracy of urban crowd-sourcing for maintaining large-scale geospatial databases. (ACM Press).

Mooney, P., and Corcoran, P. (2012). Characteristics of Heavily Edited Objects in OpenStreetMap. Future Internet *4*, 285–305.

Mooney, P., and Corcoran, P. (2014). Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors: Analysis of Interaction and Co-editing Patterns amongst OpenStreetMap Contributors. Transactions in GIS *18*, 633–659.

Neis, P., Zielstra, D., and Zipf, A. (2011). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. Future Internet *4*, 1–21.

Pell, A., Meingast, A., Rotter, S., and Pajones, M. (2013). Analysis of the applicability of the public upper austrian transport graph for solving a location allocation problem.

Petit, O., Billon, P. & Follin, J.-M. Évaluation de la qualité des données OpenStreetMap sur la Sarthe et réflexion sur le processus de contribution. XYZ 24–69 (2012).

Tenney, M. (2014). Quality Evaluations on Canadian OpenStreetMap Data.

Vandecasteele, A., and Devillers, R. (2013). Improving volunteered geographic data quality using semantic similarity measurements. (Hong Kong: International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences).

Zielstra, D., and Hochmair, H. (2012). Using Free and Proprietary Data to Compare Shortest-Path Lengths for Effective Pedestrian Routing in Street Networks. Transportation Research Record: Journal of the Transportation Research Board *2299*, 41–47.

Zielstra, D., and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. (Guimaraes, Portugal),.

Zielstra, D., Hochmair, H.H., and Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap - A United States Case Study: Assessing the Effect of Data Imports on the Completeness of OpenStreetMap. Transactions in GIS *17*, 315–334.

Zielstra, D., Hochmair, H., Neis, P., and Tonini, F. (2014). Areal Delineation of Home Regions from Contribution and Editing Patterns in OpenStreetMap. ISPRS International Journal of Geo-Information *3*, 1211–1233.

Website :

-http://openstreetmap.fr/
-http://wiki.openstreetmap.org

CE QU'IL FAUT RETENIR...

- Après un état des lieux de la bibliographie disponible sur la question de la complétude (un des principaux thèmes pour lequel OSM est cité dans la littérature scientifique), on peut retenir :
  - **La complétude et la justesse du jeu de données OSM n'est pas homogène**. Celle-ci dépend à la fois des imports officiels ayant pu être fait, du nombre de contributeurs actifs dans la zone et de leurs comportements, des besoins ou événements particuliers ayant pu conduire certains lieux à être renseignés plus efficacement dans OSM.
  - **La précision et la complétude du jeu de données semblent suivre une évolution encourageante** (identifiable notamment lors de la comparaison des résultats d'articles évaluant la qualité des données OSM paru entre 2007 et 2010 avec les articles les plus récents) et l'engouement présent autour du projet OpenStreetMap est toujours d'actualité.
  - **Les données OSM sont structurées différemment de certains autres jeux de données**, pouvant ainsi conduire à des difficultés de prise en compte de certains éléments (pourtant renseignés dans OSM), que ce soit en raison des tags utilisés ou en raison des types d'objets qu'il comporte (classés selon qu'il s'agisse de *node*, *way* ou *relation*), pouvant conduire à la perte d'information lors de la transformation des données en vue de les comparer. De plus une **diversité parfois non-désirée dans les clés/valeurs utilisées** peut résulter de l'absence d'un dictionnaire les restreignant.
- On note qu'il est possible de **visualiser les différences de rendus entre plusieurs représentations cartographiques issues d'OpenStreetMap et des représentations issues de jeux de données commerciaux** (notamment sur http://tools.geofabrik.de/mc/). Une autre application en ligne (https://mvexel.github.io/thenandnow/) permet de visualiser interactivement les différences entre le rendu en juin 2007 et le rendu actuel, mettant ainsi en avant l'enrichissement de la base de données (Figure 3).
- La question de la qualité des données est également un sujet important pour les contributeurs d'OpenStreetMap. Une page du wiki (http://wiki.openstreetmap.org/wiki/Quality_assurance) est consacrée aux **différents outils disponibles en ligne pour localiser, visualiser ou rapporter des erreurs** pouvant exister dans la base de données ainsi que pour mettre en avant des zones moins bien renseignées.
- Plusieurs outils thématiques existent (http://map.comlu.com/ pour les *turn restrictions* par exemple) parmi lesquels les plus complets, *OSM Inspector* et *Keep Right,* mettent en avant des erreurs (qu'il est possible de sélectionner parmi 13 thématiques pour *OSM Inspector* : coastline, highway, arrêts de transport en communs, etc. et les éléments comportant des clés/valeurs inhabituelles, des erreurs de connectivité ou des noms manquants) : http://tools.geofabrik.de/osmi/ et http://keepright.at.
  - Pour le territoire français il existe un moyen de visualisation des différences entre les données OSM et les données de la Base d'Adresse Nationale (BAN). Ces différences sont visibles à l'adresse : http://tile.openstreetmap.fr/~cquest/leaflet/bano.html#19/48.84356/2.37083.
  - Une plate-forme permet de mettre en avant les éléments présents dans OSM selon leur accessibilité aux fauteuils roulants et propose aux visiteurs une interface simplifiée pour enrichir OSM sur cette thématique : http://wheelmap.org/.
  - Une plate-forme permet de mettre en valeur les voies routières selon la vitesse maximum renseignée : http://www.itoworld.com/map/35.
  - Une plate-forme permet de mettre en valeur les différents éléments relatifs au transport ferroviaire : http://www.openrailwaymap.org/