ÉVALUATION DE LA QUALITÉ DE DONNÉES DE LA BASE OPENSTREETMAP PAR APPRENTISSAGE : CAS DE LA RÉPUBLIQUE DE DJIBOUTI[1]

par Ibrahim Maidaneh Abdi

Institut national de l'information géographique et forestière
73 avenue de Paris 94160 Saint-Mandé

La qualité des données de la base OpenStreetMap peut être évaluée en comparant les données d'OSM avec les données d'une base de données géographiques de référence. Cependant, en l'absence d'un organisme national de cartographie produisant une base de référence (comme c'est le cas dans la République de Djibouti), l'évaluation de la qualité des données OSM par comparaison avec une base de référence n'est pas faisable. La précision spatiale de ces données n'est alors pas connue, alors que les données OSM pourraient potentiellement servir de base à une base de données de référence. Ces travaux de thèse proposent un cadre de travail permettant d'inférer la précision spatiale des données OSM à l'aide de méthodes d'apprentissage automatique sous l'hypothèse qu'il existe une relation statistique entre une mesure de la qualité extrinsèque (précision spatiale) et des indicateurs intrinsèques de la qualité (élongation, granularité…).

La première étape est la définition d'un certain nombre d'indicateurs extrinsèques décrivant la forme et la position des objets géographiques en appariant un jeu de données OSM avec un autre jeu de données issu d'une base de référence (fig. 1). À l'issue d'un appariement multicritère basé sur la théorie des croyances, chaque objet OSM apparié se voit associer quatre valeurs mesurant des écarts par rapport à son homologue de la base de référence.

La deuxième étape est la définition d'indicateurs intrinsèques sur les bâtiments d'OSM. Ces indicateurs intrinsèques conçus comme des marqueurs de qualité de l'objet sont calculés sur les objets OSM appariés.

La troisième étape est la définition d'un modèle d'apprentissage supervisé dont l'objectif est de reconstruire les mesures extrinsèques à partir des mesures intrinsèques. Ces travaux de thèse présentent deux méthodes, une régression multiple LASSO et une classification par Random Forest. La première permet une évaluation quantitative de la précision spatiale tandis que la deuxième fournit une évaluation qualitative.

Ce framework a été appliqué sur les bâtiments à l'échelle de plusieurs communes d'un département (le Val-de-Marne). La méthode de régression appliquée à l'échelle individuelle (le bâtiment) permet d'expliquer jusqu'à 30% de la variance des mesures de précision spatiale extrinsèques, tandis que la même méthode appliquée à l'échelle agrégée (structure spatiale basée sur l'autocorrélation) permet un score de variance expliquée allant jusqu'à 42%. La méthode de classification, quant à elle, permet de détecter des bâtiments de qualité insuffisante avec une sensibilité de 80% et une spécificité de 70% (fig.2).

La transférabilité du modèle de classification est également étudiée, de sorte à mesurer la part de performance de prédiction du modèle de classification entraîné sur une zone d'étude donnée qui se conserve sur de nouvelles zones d'étude. Pour cela, deux nouvelles zones ont été considérées, la première dans le département du Gers et la seconde autour de la ville d'Édimbourg (Écosse). Cette étude montre que le modèle se transfère de manière satisfaisante.

Le classifieur de type Random Forest obtenu par apprentissage a été appliqué sur un jeu de données OSM de Djibouti pour étudier la possibilité de dériver un jeu de données de référence à partir des données OSM. 82.6% des bâtiments de la zone d'étude sont alors classifiés comme étant de qualité insuffisante, mais le modèle présente encore de nombreuses limites.

1 Thèse sous la direction d'Ana-Maria Olteanu-Raimond et soutenue à l'Université Gustave Eiffel le 23 novembre 2022

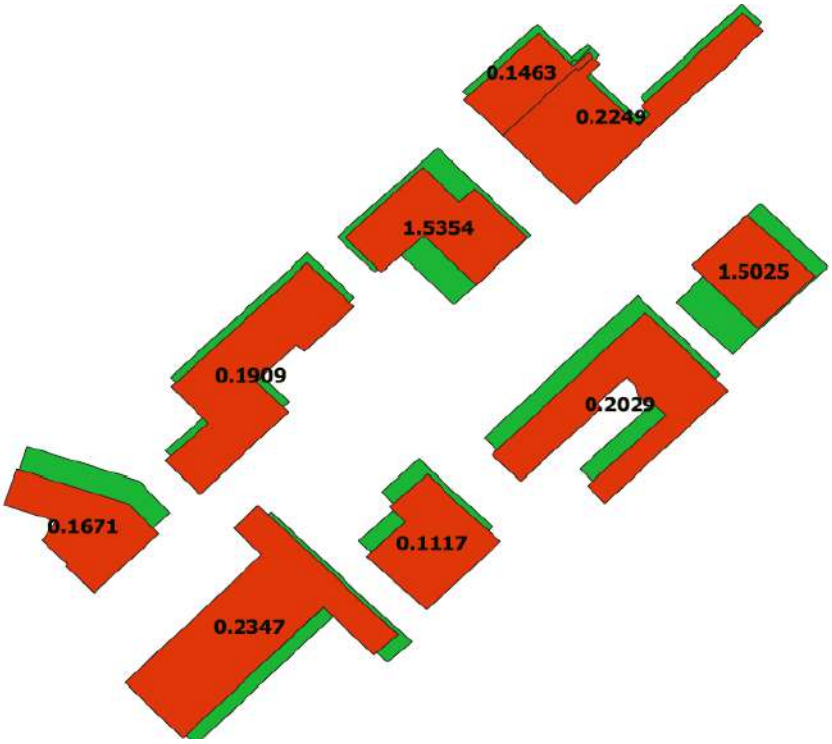

Figure 1 : Distance radiale entre bâtiments OSM et bâtiments de la base de référence candidats à l'appariement.

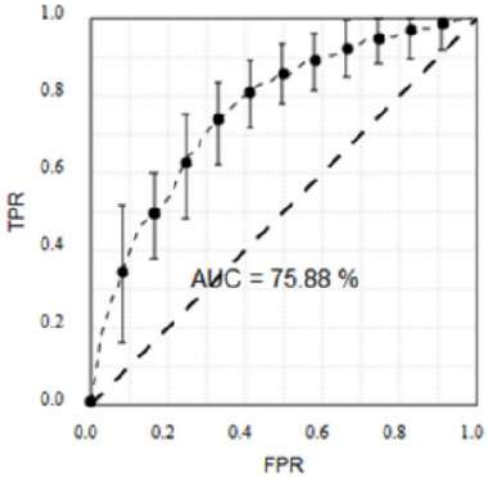

Figure 2 : Courbe ROC du classifieur Random Forest sur le jeu de données Val-de-Marne