Évaluation de la qualité des données géographiques d'OpenStreetMap à l'aide des méthodes d'apprentissage automatique : cas de la République de Djibouti

Ibrahim Maidaneh Abdi

Université Paris-Est

THÈSE

pour obtenir le grade de

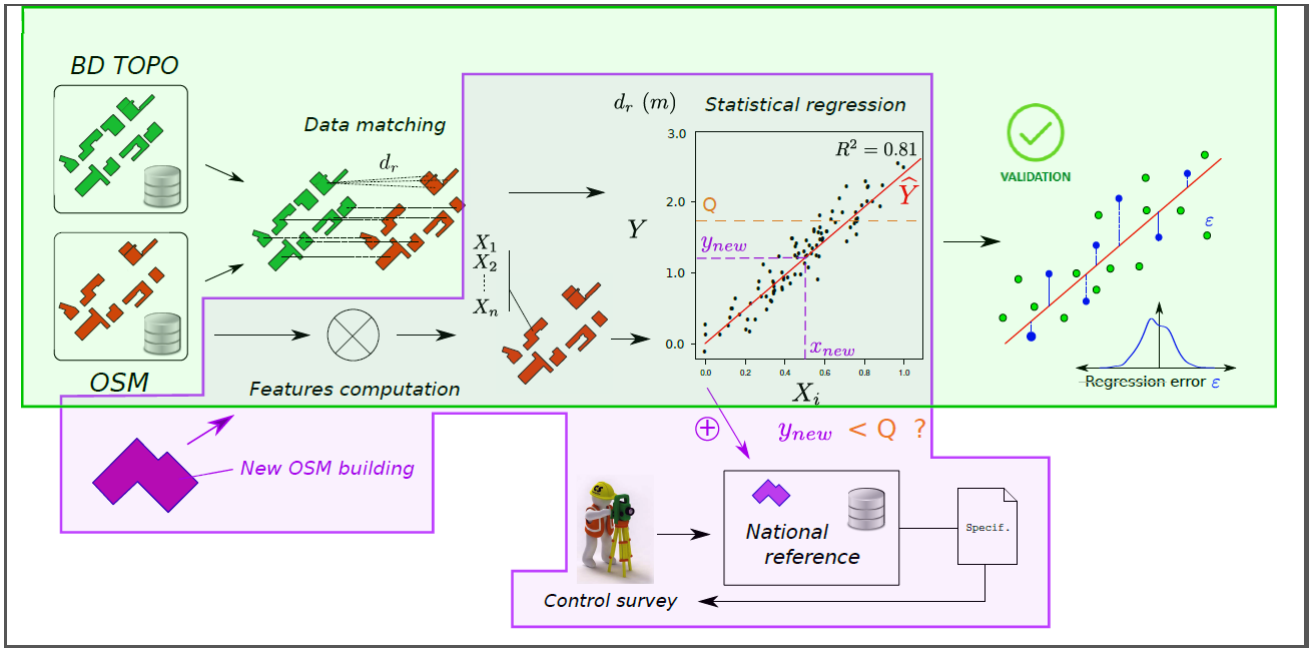
Docteur de l'Université Paris Est

Spécialité : **Sciences de l'Information géographique**

préparée dans le cadre de l'École Doctorale Mathématiques et STIC

Ibrahim MAIDANEH ABDI

Évaluation de la qualité de la base OpenStreetMap à l'aide des méthodes d'apprentissage automatique : Cas de la République de Djbouti



Rapporteure : **Anne RUAS, Ingénieur en chef des Ponts, Eaux et Forêts, HDR** (IFSTTAR)
Rapporteur : **Didier JOSSELIN, Directeur de recherche, UMR ESPACE** (Université Avignon, Côté d'Azur)
Examinateur : **Mireille Batton Hubert, Professeure** (École des Mines de Saint-Étienne)
Examinateur : **Didier JOSSELIN, Directeur de recherche, UMR ESPACE** (Université Avignon, Côté d'Azur)
Directeur de Thèse : **Ana-Maria Olteanu-Raimond, Directeur de recherche, Laboratoire LASTIG** (Univ Gustave Eiffel, IGN)
Encadrant : **Arnaud Le-Guilcher Chercheur, Laboratoire LASTIG** (Univ Gustave Eiffel, IGN)
Encadrant : **Moustapha NOUR AYEH, Maître de conférences** (Université de Djibouti)

Résumé de thèse

Généralement, 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 mettant en place une base de référence (cas de 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 pas connue alors que les données OSM pourraient servir à constituer un référentiel national géographique pour les pays ne disposant pas de base de données géographiques de référence.

Dans ce travail de thèse, nous proposons un cadre d'étude 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 (i.e. élongation, granularité).

Dans un premier temps, nous définissons et implémentons un certain nombre d'indicateurs extrinsèques décrivant la forme et la position des objets géographiques en appariant un jeu de données OSM avec un jeu de données issu d'une base de données de référence. À l'issue d'un appariement multicritères basé sur la théorie des croyances, nous parvenons à définir pour chaque objet OSM apparié, quatre caractérisant mesurant des écarts par rapport à leurs homologues de la base de référence.

En second lieu, nous définissons un ensemble d'indicateurs intrinsèques décrivant la forme et la position des bâtiments et modélisant l'erreur géométrique de saisie. Ces indicateurs intrinsèques conçus comme des marqueurs de qualité de position et/ou de forme sont calculés sur les objets OSM appariés pour disposer d'une évaluation intrinsèque de la qualité.

En troisième lieu, nous définissons un modèle d'apprentissage supervisé dont l'objectif est de décrire les mesures extrinsèques à partir des mesures intrinsèques. Nous avons proposé deux méthodes à savoir une régression multiple LASSO et une classification de type Random Forest. La première permet une évaluation quantitative de la précision spatiale tandis que la deuxième fournit une évaluation qualitative.

Ce cadre d'étude a été appliqué sur le thème bâti représentant les bâtiments à l'échelle d'un département (Val de Marne-94). 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 al-

lant jusqu'à 42%. La méthode de classification, quant à elle, permet de détecter une qualité insuffisante de bâtiments avec une sensibilité de 80% et une spécificité de 70%.

Enfin nous étudions la transférabilité du modèle de classification de sorte à mesurer la part de performance de prédiction du modèle de classification entrainé sur une zone d'étude donnée qui se conserve sur deux nouvelles zones d'étude (département GERS (France) et Ville Édimbourg (Écosse)). L'application du modèle de classification de la zone d'étude montre que le modèle se transfère de manière satisfaisante.

En guise d'application, nous avons appliqué le classifieur obtenu par l'apprentissage sur un jeu de données d'OSM de Djibouti dans le but de dériver un jeu de données de référence. Nous avons observé que 82.6% des bâtiments de la zone d'étude ont été classifié comme étant de qualité insuffisante. La comparaison des résultats avec un jeu de données de contrôle nous a permis de dégager quelques spécificités de notre modèle. Ce dernier classe en qualité insuffisante, les bâtiments ayant beaucoup d'irrégularités au niveau des côtés, les bâtiments saisis de manière trop parfaite, les bâtiments complexes regroupant plusieurs bâtiments, etc.

Generally, the quality of OpenStreetMap data can be assessed by comparing OSM data with reference spatial data. However, in the absence of a national mapping agency producing a reference database (as in the case of the Republic of Djibouti), the evaluation of the quality of OSM data by comparison with a reference database is not posible. The spatial accuracy of these data is not known, whereas OSM data could be used to constitute a national geographic reference for countries that do not have a reference spatial database.

In this thesis, we propose a framework to infer the spatial accuracy of OSM data using machine learning methods under the assumption that there is a statistical relationship between an extrinsic quality measure (spatial accuracy) and intrinsic quality indicators (i.e. elongation, granularity).

First, we define and implement a set of extrinsic indicators describing the shape and position of spatial objects (i.e. polygon) by matching OSM data with a reference spatial data. Following a multi-criteria matching based on Belief Theory, we defined for each matched OSM object, four extrinsic indicators measuring deviations from their counterparts in the reference database.

Second, we define a set of intrinsic indicators describing the shape and position of the polygones and modeling the geometric input error. These intrinsic indicators designed as position and/or shape quality markers are computed on the matched OSM objects to have an intrinsic quality assessment.

Third, we define a supervised learning model whose objective is to describe the extrinsic measures from the intrinsic measures. We proposed two methods, namely a LASSO multiple regression and a Random Forest classification. The first one allows a quantitative evaluation of the spatial accuracy while the second one provides a qualitative evaluation.

The proposed framework was applied to the building theme representing buildings at the scale of a department (Val de Marne-94). The regression method applied to the individual scale (i.e. building) allows to explain up to 30% of the variance of the extrinsic

spatial accuracy measures, while the same method applied to the aggregated scale (spatial structure based on autocorrelation) allows an explained variance score up to 42%. The classification method, on the other hand, detects insufficient building quality with a sensitivity of 80% and a specificity of 70%.

Finally, we study the transferability of the classification model in order to measure the part of prediction performance of the trained classification model on a given study area which is maintained on two new study areas (GERS department in France) and Edinburgh city (Scotland)). The application of the classification model to the study area shows that the model transfers well.

As an application, we applied the classifier obtained by the training on an OSM dataset from Djibouti in order to derive a reference dataset. We observed that 82.6% of the buildings in the study area were classified as poor quality. The comparison of the results with a control dataset allowed us to identify some specificities of our model. The latter classifies as poor quality, the buildings having many irregularities at the level of the sides, the buildings captured in a too perfect way, the complex structures grouping several complex buildings, etc.

Table des matières

Chapitre 1

Introduction générale

Auparavant, en cartographie numérique, les organismes étatiques développaient des systèmes d'information géographique permettant à chacun d'entre eux de gérer, d'analyser et de diffuser ses propres données en interne. On peut définir une rupture au moment au moment où ces organismes fournissent au grand public un accès à leurs données relatives aux infrastructures, à l'occupation du sol, à l'urbanisme, etc. et cela à travers le développement des outils de Web SIG. À travers des interfaces cartographiques, les utilisateurs peuvent manipuler les données géographiques avec le traitement et l'affichage de leur choix sans aucune restriction d'usage à l'égard des différentes thématiques mises à disposition.

Plus tard, une nouvelle rupture est causée par l'émergence de ce que l'on appelle le Géoweb (ou le Geospatial web) défini comme une organisation de l'information d'Internet à travers un géoréférencement direct ou indirect sur la surface terrestre de tout contenu informationnel (Joliveau, 2011). Le Géoweb offre aux utilisateurs des applications géographiques composées des services cartographiques et des contenus Web en s'appuyant sur des API (Application Programing Interface) comme celles de Google Maps, Bing Maps, ESRI ArcGIS ou du Géoportail de l'IGN. Ces applications assimilées à des portails Web permettront à des utilisateurs ayant peu de connaissances techniques de s'approprier des fonctionnalités d'édition, de création, de diffusion des données géographiques et parfois d'échange avec d'autres internautes afin d'assurer une production continuelle de contenu, et non plus uniquement de la diffusion de la donnée géographique. Ce phénomène se qualifie de cartographie 2.0 (Joliveau et al., 2013). La mobilisation de ces technologies géospatiales impliquant une grande interactivité du grand public crée de grandes quantités de données générées par les utilisateurs. Ces technologies ajoutées à des objectifs de collecte, de création commune et de partage de données géographiques, donnent naissance à des mouvements collaboratifs qui aboutiront à une carte d'information géographique dite la **carte collaborative** (Mericskay et Stéphane, 2010).

Afin de mieux cerner la cartographie collaborative, nous catégorisons plusieurs types d'information géographique donnant plusieurs approches de la cartographie, en fonction de leur producteur. Dans un premier temps, nous disposons de l'information géographique institutionnelle détenue par les agences publiques, comme les services nationaux qui dressent les cartes générales (par exemple l'IGN, Institut national de l'information géographique et forestière en France). Ces agences ont comme mission principale de produire des bases de données de référence, qui parfois mettent en ligne les référentiels issus de ces bases de données dont la diffusion peut être libre/gratuite ou bien payante. Dans ce contexte

d'information géographique dite d'autorité, l'utilisateur n'intervient à aucun moment dans sa production (Mericskay, 2008). Un autre type d'information géographique est l'information géographique commerciale créée par des entreprises privées. Elle se présente sous la forme de POI (Points d'intérêt) ou de réseau routier, visualisés sur des cartes en ligne (Mericskay et Roche, 2011). L'information commerciale peut parfois être disponible gratuitement sur certaines applications notamment les applications de calculs d'itinéraires. En parallèle à ces deux formes classiques de l'information géographique, s'est développée une information géographique qualifiée de volontaire (Goodchild, 2007) donnant naissance à la cartographie collaborative. Étant associée à la cartographie du Web 2.0, la cartographie collaborative se fonde sur un internet participatif dans lequel les volontaires créent de l'information géographique soit très spécifique, en participant à une volonté commune comme une carte d'édition de toutes les pistes cyclables urbaines de la région, ou plus générique, par exemple une carte du monde placée sous licence libre.

Sous un angle applicatif, quand la cartographie a été en particulier mise en œuvre dans le cadre de projets de développement des territoires, supposant une implication directe et non virtuelle de leurs habitants (autour de l'urbanisme, en aménagement) afin de créer un produit collectif issu d'un projet de collecte de données (sur une thématique donnée), on parle de cartographie participative (Palsky, 2013). Le citoyen est amené à contribuer de chez soi, sans compétences particulières et à apporter ainsi son expertise : on parle souvent de l'expert amateur. L'initiative d'impliquer le citoyen dans la cartographie de son environnement proviendrait d'un manque d'information géographique localement de la part des décideurs territoriaux ou bien du souhait des politiques publiques d'impliquer les citoyens. Elle s'organise autour des applications SIG web à travers lesquelles les citoyens sont censés contribuer de la donnée géographique.

Par ailleurs, il existe un autre type de cartographe collaborative (Lambert et Zanin, 2012) dont le but final est de créer une cartographie libre du monde dont chaque producteur est à la fois contributeur et utilisateur de l'information géographique et où une certaine qualité est assurée par le phénomène du crowdsourcing (approvisionnement par la foule) (Haklay et al., 2010) correspondant à la capacité d'une communauté de contributeurs à mettre à jour des données et à corriger les erreurs par des contributions individuelles d'un grand nombre de participants. L'un des projets issu du mécanisme d'édition de l'information géographique volontaire (VGI) que l'on notera initiative VGI est celui d'OpenStreetMap (OSM).

Un des points forts d'OpenStreetMap avec certains des outils de cartographie collaborative, est que tous les contributeurs d'OpenStreetMap travaillent sur la même carte. Cela garantit que les données suivent le même format partout dans le monde tout en permettant aux contributeurs d'ajouter de plus en plus de détails à partir de nombreuses sources. Cette démarche empêche l'information d'être divisée et bloquée dans des projets individuels. De plus, OpenStreetMap dispose d'une grande communauté active. Il a été recensé qu'en France **200 à 300** comptes actifs par jour mettent à jour la base OSM [1] bien que ces comptes actifs ne constituent que 10% des contributeurs d'OSM avec 1% de ces contributeurs ayant produit 50% des contributions (d'après une enquête menée à Bordeaux par Noucher (2014)). Cette communauté dynamique dans son ensemble comprend les personnes qui ajoutent des informations à la carte, celles qui écrivent des logiciels pour permettre la modification et d'autres qui créent des cartes spécialisées à partir des données

1. La conférence State of the Map France en 2016

OpenStreetMap. Les adeptes d'OpenStreetMap se réunissent partout dans le monde pour partager des techniques et cartographier leurs lieux de vie.

Beaucoup d'applications se basent sur la cartographie OSM pour constituer une cartographie partagée entre des adeptes d'une même cause. On cite en guise d'exemple, une application web portée par les utilisateurs de vélo, d'éco-mobilité en général et de cartographie, la carte **Opencyclemap** [2] qui rassemble les pistes cyclables nationales ou urbaines ainsi que d'autres informations utiles aux cyclistes. Concernant l'assainissement, l'accès à l'eau potable, les cheminements et la structure de l'habitat, les données cartographiques libres complètent les efforts des services étatiques et des organisations humanitaires dans les pays en développement afin de comprendre les conditions de vie et la structure de l'habitat des gens qui vivent dans de vastes bidonvilles.

Du fait de la richesse de l'information géographique présente dans OSM, l'évaluation de la qualité des données d'OpenStreetMap est une question importante. L'évaluation de la qualité d'un jeu de données OSM porte sur plusieurs critères, à savoir principalement la précision de la localisation des données, la complétude brute des donnée, la complétude et la justesse des attributs (Senaratne et al., 2017; Mooney et Corcoran, 2012a,a; Koukoletsos et al., 2012; Neis et al., 2012), l'historique des données et enfin l'évaluation de la crédibilité des contributeurs (Bishr et Janowicz, 2010; Lodigiani et Melchiori, 2016).

La Littérature distingue deux manières de qualification des données collaboratives à savoir une évaluation extrinsèque et une évaluation intrinsèque. Dans le cas d'une évaluation extrinsèque de la qualité, les chercheurs peuvent comparer un jeu de données OSM avec un autre de jeu de données issu d'une base de référence en utilisant un certain nombre de mesures de la qualité (précision géométrique, précision thématique et attributaire, etc.) (Kounadi, 2009; Ather, 2009; Helbich et al., 2012; Fan et al., 2014; Girres et Touya, 2010; Graser et al., 2014), soit en menant une évaluation de la crédibilité des contributeurs à travers des indicateurs de qualité sur le contributeur (confiance, expertise, connaissance locale, etc.) (Flanagin et Metzger, 2000; Begin et al., 2016; Coleman et al., 2009; Kuai et al., 2016).

Dans le cas d'une évaluation intrinsèque de la qualité, les études de recherche définissent des descripteurs relevant des caractéristiques sur la forme et sur la position des objets d'un jeu de données OSM de sorte à quantifier ou classifier une qualité de saisie de ces objets (granularité, convexité, compacité, élongation, nombre de sommets par rapport au périmètre, la cohérence topologique, etc.) (Ivanovic et al., 2019; Ciepłuch et al., 2011; Antoniou et Skopeliti, 2015). Quant aux études menant une évaluation intrinsèque en se basant sur l'historique des données, les auteurs évaluent la qualité à travers entre autres le nombre d'édition par entité géographique (Keßler et al., 2011), les incohérences topologiques (Hashemi et Abbaspour, 2015; Touya et Brando-Escobar, 2013), le nombre moyen de tags par entité géographique, le nombre total des contributeurs actifs sur l'entité géographique, etc. (Barron et al., 2014). Enfin, en se fondant sur le contexte spatial, des études mesurent la qualité intrinsèque à travers la cohérence du thème d'un objet géographique avec sa localisation, la comparaison entre deux thèmes d'un même objet, etc. (Jolivet et Olteanu-Raimond, 2017).

Dans les travaux de recherche sur OSM, les chercheurs observent une hétérogénéité

2. http ://www.opencyclemap.org

spatiale des données OSM (Yang et al., 2018; Minaei, 2020) mais en même temps ils attestent que les données OSM ont une bonne qualité en précision spatiale et en complétude (Ma et al., 2015; Viana et al., 2019) au moins en zones urbaines densément peuplées. L'importation dans OSM de données issues de bases de référence contribue à la précision et à la complétude de la base de données OSM. Sur ces études, ils constatent que la majorité d'entre elles a porté sur la précision de position des objets ponctuels et la complétude d'objets linéaires comme le réseau routier. Enfin, ils estiment que comportement des contributeurs conditionne la fiabilité des données OSM (Senaratne et al., 2017).

Face à ces considérations sur l'évaluation de la qualité des données OSM, notre questionnement scientifique de thèse se positionne dans un contexte, à Djibouti, d'absence de données de référence comme dans la plupart des pays d'Afrique. Ce contexte se caractérise par un absence d'organisme national de cartographie qui produit des données géographiques de référence, par une absence de spécifications pour produire des données géographiques. Cela engendre des lacunes dans les modèles de prise de décision. Une solution consisterait alors à utiliser les données collaboratives comme données de référence après les avoir qualifiées.

Ainsi, la problématique de recherche de notre thèse concerne l'évaluation de la qualité des données OSM en absence de données géographiques de référence, à l'aide des méthodes d'apprentissages.

Pour cela, nous fixons un objectif de thèse consistant à proposer un cadre d'étude nous permettant d'inférer la précision spatiale des données OSM à l'aide de méthodes d'apprentissage automatique sous l'hypothèse qu'il existe un lien statistique entre la précision spatiale des objets géographiques et des indicateurs intrinsèques pouvant caractériser la qualité de saisie.

Le manuscrit de thèse est organisée comme suit.

Le chapitre 2 présente le cadre général et la problématique de la thèse en commençant par une introduction concernant les aspects de la qualité, en présentant les initiatives des projets VGI et en analysant les potentialités d'utilisation des données OSM, pour finir par détailler le contexte menant à la formulation de notre problématique de thèse et les objectifs qui en découlent.

Dans le chapitre 3 nous réalisons une synthèse de l'état de l'art sur la question centrale de notre thèse à savoir la qualification des données géographiques volontaires en passant par une présentation du cade théorique sur la qualité des données géographiques puis des démarches d'évaluation de la qualité pour finir par une synthèse d'état de l'art sur les outils utilisés dans le processus de qualification des données géographiques que nous nous apprêtons à déployer.

Dans la seconde partie de notre manuscrit de démarche et mise en œuvre, en chapitre 4, nous décrivons d'abord de manière générique, notre cadre de travail méthodologique qui guidera tous nos travaux d'expérimentation, d'implémentation et de réalisation visant à mener une évaluation intrinsèque de la qualité.

En chapitre 5, nous examinons l'ensemble des algorithmes portant sur l'appariement

des données géographiques pour se choisir et réaliser un processus d'appariement adéquat à notre problématique et aux données OSM.

Dans le chapitre 6, nous dressons notre modèle d'apprentissage d'inférence de la qualité extrinsèque sur deux méthodes à savoir une régression multiple LASSO et une classification de type Random Forest.

En chapitre 7, la discussion sur nos résultats nous mène à étudier la transférabilité du modèle d'apprentissage proposé. Après avoir validé la généralisation de notre modèle sur des nouvelles zones d'études, nous détaillons les résultats obtenus du modèle de classification sur un jeu de données d'OSM de Djibouti.

Première partie

Contexte applicatif et scientifique de la thèse

Chapitre 2

Cadre général et problématique

Sommaire

Il est estimé que plus de 85% des bases de données ont une composante géographique [1]. Quand une information est localisée dans l'espace avec une certaine représentation (i.e. point, ligne, surface), on parle d'une information géographique. Le cycle de vie de l'information géographique est composé de quatre grandes phases. Pour pouvoir produire de l'information géographique, il est nécessaire de se positionner par rapport à un système de cordonnées dans l'espace, c'est la phase du géoréférencement. Puis arrive la phase d'acquisition des mesures qui sont effectuées à l'aide d'instruments et de techniques, et leur transformation en objet géographique. Ensuite, on passe à la une phase de traitement des données dont l'objectif principal est de représenter et structurer les données géographiques pour produire une information géographique compréhensible et exploitable. Enfin, l'information géographique est stockée dans des bases de données géographiques parfois exploitables sur le Web et peut être diffusée soit sous format vecteur soit sous forme des cartes.

Durant ce cycle, on produit des données dites traditionnelles pour lesquelles les techniques d'acquisition, les appareils de mesure ainsi que la précision attendue sont fixés, y compris l'expertise du technicien qui manipule les appareils et qui produit la donnée

1. https ://calenda.org/209584

géographique.

Avec l'avènement du Géoweb, il est devenu possible pour le grand public de localiser une information géographique en ligne. Des bases de données éditables en ligne naissent d'une volonté de créer un bien commun de la connaissance de la donnée géographique, volonté autour de laquelle se dessine une communauté issue en particulier de la mouvance du Libre (open source, open data, open accès, open culture, open hardware, etc.) et ayant un centre d'intérêt commun : *cartographier le monde librement*. C'est dans ce cadre qu'émergent des projets collaboratifs.

Les projets collaboratifs s'insèrent dans ce qu'on appelle de contenus géographiques générés par les utilisateurs (Purves et al., 2011), de crowdsourcing géographique (Sui et al., 2013), de néogéographie (Turner, 2006) et du Géoweb (Leszczynski et Wilson, 2013) respectivement selon la communauté de la recherche, de la géographie ou de l'informatique, de la cartographie ou de la géomatique et la communauté de cartographie ou de géomatique. Toutefois, Goodchild (2007) propose d'utiliser la notion de *Volunteered Geographic Information* (VGI) comme un terme permettant de réunir sous un même vocable l'ensemble des démarches de création de contenus géolocalisés, bénévoles et spontanées qui fournissent aujourd'hui des données géographiques des différentes des productions conventionnelles des professionnels du secteur.

Si les données VGI ont en commun avec les données traditionnelles de proposer à la fois une information géométrique et sémantique, elles diffèrent en revanche de par leurs méthodes d'acquisition.

Une analyse comparative entre les données traditionnelles et celles des projets volontaires serait utile pour mieux comprendre et exploiter ces données. Cette analyse tente de répondre à la question suivante en termes de temps et de techniques mises en œuvre et d'autres considérations externes :

- En quoi l'acquisition de données VGI diffère-t-elle de la collecte des données traditionnelles ?
  Après une brève définition de la qualité de l'information géographique et les démarches qualité, on analyse la collecte des données traditionnelles et des VGI pour en tirer une comparaison révélant les forces et les faiblesses des données volontaires.

- Quels usages peut-on dresser autour des données géographiques volontaires ?
  Nous examinons ensuite les différents projets VGI de sorte à les définir et à les classer dans leur contexte d'utilisation. Par la suite, nous nous concentrerons sur les données de la plateforme OpenStreetMap (OSM) pour analyser les potentialités d'usages des données OSM qui pourraient s'insérer dans un contexte précis tout en s'alignant aux défis de la recherche sur la qualification des données.

2.1 Introduction

2.1.1 Définition de la qualité des données géographiques

Dans la littérature, la donnée est souvent intégrée dans une hiérarchisation avec les concepts d'*information* et de *connaissance*. Les chercheurs distinguent deux types de

hiérarchie selon le mode ascendant ou descendant.

Selon la vue descendante, certains chercheurs considèrent que *les données* sont des nombres et des faits bruts ; les informations sont des données traitées et la connaissance une information authentifiée, possédée dans l'esprit des individus : c'est une information personnalisée (qui peut ou non être nouvelle, unique, utile ou exacte) liée aux faits, procédures, concepts, interprétations, idées, observations, et jugements (Dretske, 1981; Machlup et Mansfield, 1983; Vance, 1997). De ce principe, nous procédons à l'élaboration de l'information à partir de données. Puis cette information devient admise, valide et appropriée pour chacun pour en faire une connaissance.

Selon la vue ascendante, Tuomi (1999) postule que la connaissance doit exister avant que l'information puisse être formulée et que les données puissent être mesurées pour former des informations. L'auteur nie l'existence en soit de *données brutes* et affirme que même la partie la plus élémentaire de *données* a été influencée par les processus de réflexion ou de connaissance qui ont mené à son identification et à sa collecte. Cela suppose de considérer une connaissance, puis de fixer la façon de déterminer la donnée et de produire une information à partir de cette donnée.

Ayant considéré ces deux manières de situer une *donnée*, nous définissons tout d'abord qu'une donnée est une observation, une description ou encore une mesure faite sur des objets du monde réel de manière à mesurer soit leur position (donnée géographique) soit une caractéristique (donnée sémantique). Cette mesure de la réalité est précédée par une conceptualisation ou modélisation du monde réel qui définit la manière d'effectuer cette mesure. L'information géographique est produite lorsqu'on associe cette donnée géographique à une description sémantique ou attributaire définissant la donnée géographique dans son contexte précis. Elle se matérialise en adoptant une représentation de la donnée, qui peut être obtenue immédiatement ou après l'application de traitements. Lorsque ce modèle de mesure et cette représentation sont communément acceptés par une communauté scientifique, on parle d'une *connaissance*.

Au sens du management de la qualité, selon la norme ISO 9000 :2000 [ISO 00], la qualité correspond à toutes les caractéristiques d'un produit propres à satisfaire des exigences. Nous avons d'une part des caractéristiques qui constituent les *paramètres de la qualité* et d'autre part des exigences qui correspondent aux conditions de production, le lien affirmant la validité, la qualité étant la satisfaction ou la conformité.

Du point de vue géographique, la qualité de données géographiques doit faire attention à la notion de la *représentation*. Cette dernière est une abstraction selon un modèle de signes de ce qui doit être mesuré. Ce qu'on appelle sous une autre forme la *dénotation* (Hangouët, 2005). Dans notre cas, cette notion de représentation fait aussi appel à la notion de *terrain nominal*.

Un terrain nominal est une représentation idéale du monde réel ou d'un phénomène géographique et constitue idéalement *ce que devraient être les données produites*. Ceci est clarifié ou instancié vis-à-vis des types des objets à représenter dans ce qu'on nomme *les spécifications*. Ainsi la production d'une donnée géographique doit se faire à travers un modèle d'opérations à effectuer et sa mise en œuvre (Hangouët, 2005).

Le modèle des opérations à effectuer précise l'ensemble des opérations ou de mesure à faire depuis la technologie des capteurs à employer (la précision de l'appareil) jusqu'aux méthodes utilisés (indicateurs, statistiques calculées, transformations, classifications, pour mettre les résultats aux formats souhaités (Hangouët, 2005).

Ainsi la qualité de données géographiques pourrait être définie comme étant la fidélité ou la conformité des données produites à des spécifications (au modèle des opérations et leur réalisation, en faisant illusion à l'erreur des appareils de mesure et aux méthodes mathématiques utilisées). Cette conformité est évaluée par rapport à une ou plusieurs caractéristiques des données géographiques (les paramètres ou critères de la qualité définis par les normes ISO [2]). La conformité des données aux spécifications correspond à la qualité interne tandis que la qualité externe repose sur l'adéquation aux besoins des utlisateurs (Devillers et al., 2007).

2.1.2 Démarches pour l'évaluation de la qualité des données géographiques

Afin de garantir une qualité de l'information géographique, il faut adopter des démarches qualité. Tout d'abord, le producteur doit mettre en place une démarche d'assurance qualité qui s'opère lors de la collecte de données. L'assurance qualité peut être définie selon l'ISO 8402 [3] : *Ensemble des actions préétablies et systématiques réalisées au fur et à mesure de la production et qui seront nécessaires pour donner la confiance en ce qu'un produit satisfera aux exigences données relatives à la qualité.* Le but est ici de mettre en place un ensemble de méthodes et de contrôles tout au long du processus d'acquisition (saisie, interpolation, modélisation, etc.) afin d'obtenir des données répondant aux exigences de la qualité et pour lesquelles on pourra fournir les valeurs des composantes de la qualité (Ubeda, 1997).

Puis, il faut adopter une seconde démarche de *contrôle de qualité* afin d'évaluer ou de confirmer la qualité d'un jeu de données. Il s'agit de mettre en place des calculs qui permettront d'établir l'aptitude des données à répondre aux exigences à travers le contrôle de qualité dont l'ISO 8402 nous donne la définition suivante : *Ensemble des actions de mesure, d'examen, d'essai, de calibrage d'une ou plusieurs caractéristiques d'un produit et de comparaison aux exigences spécifiées en vue d'établir leur conformité.*

Enfin, une dernière technique de traitement concernant la qualité, peu abordée dans la littérature, peut être définie comme *l'amélioration de la qualité.* N'ayant pas été définie de manière officielle dans la littérature existante,*l'amélioration de la qualité* consiste en un ensemble de méthodes visant à corriger les erreurs contenues dans les jeux de données. Une erreur est définie vis-à-vis d'une composante de la qualité, comme un trop grand écart entre une donnée et son équivalent sur le terrain nominal (Ubeda, 1997).

Pour pouvoir évaluer la qualité d'un jeu de données nous faisons appel à un jeu de données de *référence* comme l'ont souligné David et Fasquel (1997). Ces données de contrôle sont définies comme *l'ensemble de données qui permettent de mesurer ou de contrôler la qualité d'un jeu de données par comparaison avec un échantillon du jeu de données à évaluer.* Les données de référence, qui sont donc utilisées pour réaliser des com-

2. https ://www.iso.org/fr/standard/32575.html
3. https ://www.iso.org/fr/standard/20115.html

paraisons, ont pour objectif de représenter le terrain nominal. En effet, le terrain nominal est une notion abstraite, commode à présenter de façon théorique mais qui n'est pas accessible dans la réalité. Les données de référence permettent donc de fournir une estimation *proche* (une approximation) du terrain nominal. Pour pouvoir réaliser des comparaisons entre un jeu de données à évaluer et les données de référence, un échantillon d'objets représentatifs doit être extrait. Chaque objet de l'échantillon doit ensuite être mis en correspondance avec son homologue présent dans les données de référence. Cette phase dite d'appariement de données, a fait l'objet de nombreux travaux de recherche pour faciliter son automatisation (Mustière et Devogele, 2008; Raimond et Mustière, 2008).

2.2 Comparaison des techniques de production des données géographiques

2.2.1 Techniques traditionnelles d'acquisition des données

Les premières techniques d'acquisition des données géographiques furent des mesures topométriques qui consistaient à mesurer des distances. En partant du constat qu'il est plus aisé de mesurer des angles que les distances, une meilleure stratégie consista à mettre en place un réseau de triangulation globale, robuste, raffiné au second ordre et au 3ème ordre. Pour déterminer les coordonnées d'un point, on se base sur un triangle formé entre ce point et deux autres points de coordonnées connus. A l'aide d'un appareil de topographie, on effectue des mesures de direction depuis les deux autres points, desquelles on déduit deux angles pour calculer les coordonnés du point recherché. Toutefois des erreurs sont susceptibles de se cumuler d'un point à un autre. C'est pourquoi, il est nécessaire de renforcer le réseau à l'aide des mesures de longueurs et des points par GNSS (ligne de base). On répartit les erreurs par une compensation totale par moindre carrées. L'ensemble des points obtenus constitue un canevas topographique permettant le relevé des points d'intérêts par des procédés dits de rayonnement et de cheminement. Le nivellement, quant à lui, permet de mesurer le dénivelé entre deux points, ce qui permet de calculer l'altitude d'un point à partir de celle d'un autre.

Une mission de levés de terrain par topométrie nécessite une expertise humaine à pouvoir manipuler les appareils topographiques (mise en place du trépied et du niveau, calage à la verticale). Il faut également retrouver les balises (bornes géodésiques de coordonnées connues) sur lesquelles s'appuieront les mesures. Néanmoins, des erreurs liées à la mire, à l'atmosphère, à la sphéricité de la terre, au milieu ou la dénivelée imposent des corrections. La précision d'un relevé topométrique est de l'ordre de 1 cm mais la portée des appareils est limitée à 5 km. Pour couvrir une grande surface, il est nécessaire d'effectuer plusieurs déplacements, et il faut reprendre la mise en place à chaque déplacement (stationnement, visée de points de référence...etc.).

Par ailleurs, une autre technique d'acquisition de données géographiques est celle de prise de vues aériennes (prises à la verticale) se chevauchant pour permettre une restitution par stéréoscopie. La stéréoscopie cherche à reproduire le relief à l'aide de deux images prises dans deux angles différents d'une même zone. La qualité de la perception du relief dépend du rapport Base sur Hauteur. La Base correspond à la distance (en centimètre) d'un sommet sur deux clichés successifs (images) tandis que la Hauteur correspond à la hauteur de vol du vecteur (l'appareil porteur de la caméra/appareil photo). Quand la hauteur de vol est très élevée, le rapport est assez défavorable, c'est le cas des prises de

vue par satellite. Avec un rapport moyen, on opère dans le domaine de la photographie aérienne faite par des avions ou des hélicoptères voire un drone selon les difficultés d'accès à l'environnement à cartographier. Enfin pour un meilleur rapport de Base-Hauteur, et que l'équivalent de la hauteur étant ici la distance de la caméra à la scène, on utilise des caméras au sol afin de relever tout objet non visible depuis le ciel (façades, intérieurs) à condition de déplacer les caméras afin d'assurer la stéréoscopie. En photogrammétrie architecturale, la résolution est de l'ordre de 1 cm tandis qu'en photogrammétrie aérienne elle s'élève de l'ordre de 1 m(modèle numérique de données). Quant à la photogrammétrie spatiale, la résolution est de l'ordre de 1 km.

Une prise de vues nécessite une longue phase de planification dite stéréo-préparation et des travaux topographiques connexes. Une étape de stéréo-préparation correspond à la préparation des appareils notamment la correction de la distorsion de la caméra et de la réfraction à l'entrée. La campagne de prise de vues a également besoin d'un réseau des points d'appui (pour déterminer l'orientation externe des clichés) qui doivent être visibles depuis le ciel. Après la prise des images, on procède aussi à une étape de saisie de points de liaison (orientation relative entre les clichés). Elle coûte assez cher et sa tenue dépend fortement des conditions météorologiques. Tout ceci fait que la mise à jour des prises de vues demeure assez lente. Pour l'IGN, elle se réalise tous les trois ans en plus d'un temps de traitement à la restitution photogrammétrie.

Cependant, cette technique d'acquisition d'images a ses vertus. La prise de vues peut couvrir des grandes zones et peut fournir des produits d'images assez variés et nombreux notamment la production des MNT, des MNS, une couverture sur le thème Bâti ainsi que sur les forêts. Néanmoins, les erreurs demeurent fortement corrélées sur ces produits.

Une autre technologie, la technologie LiDAR (Laser Imaging Détection and Ranging) est une technique d'acquisition assez robuste basée sur le laser. Le principe consiste à envoyer une onde sur une surface puis mesurer le temps que met le laser (l'onde) à revenir à la source. Cela permet de mesurer la distance du point d'impact. Le laser détermine également la réflectance du point d'impact afin de lui attribuer une couleur estimative. Le laser fixé à une tête rotative, il assure un balayage sous forme d'une sphère solide ou d'une couverture panoramique de la zone.

La résolution d'acquisition dépend de la résolution angulaire et de la distance de la cible si bien que la densité des points collectés dépend fortement du contexte d'utilisation. La position relative entre les points est aussi très précise à condition de bien connaitre la trajectographie de la plateforme. Ainsi la résolution du LiDAR varie entre 50 cm à 10 m selon les applications allant du scanner terrestre vers le LiDAR aéroporté.

La technologie LiDAR parait assez simple à utiliser du fait qu'elle n'exige pas les mêmes traitements que les images aériennes mais elle fournit un nuages des points massifs et non structurés qui rendraient difficile le passage à un produit de cartographie fine. Du fait de la cherté des appareils LiDAR, cette technologie s'utilise pour des observations très locales et appartient souvent à des entreprises privées. Un exemple d'utilisation du laser est le produit Litto3D publié sur le site du Géoportail de l'IGN. Il s'agit d'un modèle numérique continu terre-mer sur la frange littorale.

En plus des technologies précédentes, la numérisation manuelle ou automatique se fait

soit à l'aide d'un appareil de stéréo-restitution soit par la digitalisation sur une image par un opérateur (au bureau) à l'aide des lunettes 3D. Ce procédé induit des erreurs dûes à des inattentions de l'opérateur et de ce fait, il s'avère difficile de réaliser une uniformisation des erreurs entre opérateurs si bien qu'un contrôle de qualité s'impose. Ainsi, la mise à jour des données avec cette technique est lente et parfois coûteuse compte tenu de certains détails inaccessibles. Par exemple des enclosions en ville réclament l'obtention d'images supplémentaires et ralentissent le travail humain, parce qu'elles rendent la restitution impossible.

Enfin, une dernière technique d'acquisition traditionnelle des données géographiques consiste à simplifier des formes des objets existants après avoir sélectionné certains objets à conserver du fait d'une réduction d'échelle sur carte donnée. Il s'agit un processus de généralisation. Cette opération s'insère dans un champ de recherche actif basé sur des connaissances d'infographie et des expériences psychologiques (Ruas, 1999).

2.2.2 Techniques d'acquisition des données géographiques volontaires

La création de la donnée géographique volontaire fait allusion à tout projet de création de base de données géographiques, dont les producteurs seraient de contributeurs anonymes mais actifs. Dans un espace de communauté collaborative, la création d'un contenu partagé dans un réseau fait que producteur et consommateur de l'information géographique, soient confondus. On parle de *Producer* (Bruns, 2008). Les producteurs sont libres et même appelés à participer à la création de la connaissance (caractère libre), autorisés à l'utiliser dans leurs applications (le caractère ouvert) et à diffuser la connaissance (caractère partage). Suite à une production massive de données dans le web le terme VGI se voit complété par l'expression *contenus générés par des utilisateurs* (CGU) et enfin le terme *citoyen capteur* (Foody et al., 2017). Bien que l'acronyme CGU renvoie surtout aux tweets et aux images geotaggées sur Flicker volontairement publiés, l'information géographique qui leur est associée n'est pas volontaire tandis que le terme citoyen capteur s'insère dans un cadre de "communauté" visant à la réalisation d'un projet de collecte commun portant sur un thème précis, d'où le terme "collaboratif". Par conséquent, l'acquisition de données géographiques volontaires s'opère par deux manières de création, soit explicite ou soit implicite.

Quand un contributeur mène une activité dont le but principal consiste à cartographier les éléments de son entourage, ou un évènement à caractère géographique, on parle de création d'information géographique explicite, pour laquelle les principaux projets sont OpenStreetMap, Wikimapia tandis que la création de la donnée géographique implicite peut être issue de toute forme de donnée dans le web qui faire référence à une localisation que ce soit une image, un texte ou une vidéo, notamment les articles de Wikipedia, des images geotaggées ou encore des commentaires sur TripAdvisor qui restent des données partiellement involontaires. Par exemple une image peut être géolocalisée à travers les coordonnées du lieu de capture, et de ce fait pourrait nous informer sur l'apparition des nouveaux détails géographiques (construction d'un nouveau bâtiment, ou de l'ouverture d'une nouvelle route) (Antoniou et al., 2010; Craglia et al., 2012).

Antoniou et al. (2010) nous fournissent un panorama des technologies d'acquisition de données géographiques volontaires, tout en soutenant que ces technologies ont fait émerger le terme de *néogéographie* (introduit par Turner (2006)), un néologisme décrivant la parti-

cipation croissante des amateurs à la production des données cartographiques. Ce sont les interfaces de programmation d'application (API) en JavaScript, les appareils quotidiens compatibles avec le GPS et la technologie AJAX. Les auteurs estiment que ces technologies ont ouvert la voie à la collecte massive, la gestion et la diffusion des VGI, réduisant les coûts élevés de la production de données géographiques.

En effet, avec l'avènement de ces technologies, et du fait de la disponibilité d'une grosse quantité d'information géographique dans le Web, l'utilisateur passe d'une stade de consultation à un stade d'interaction (en l'occurrence de modification ou production) avec l'information géographique, ce qui introduit le concept des contenus générés par les utilisateurs (Goodchild, 2007) et en même temps le terme Géoweb 2.0 (Haklay et al., 2008).

D'abord, dans la cadre de la cartographie explicite la production de l'information géographique volontaire s'effectue sur des applications de cartographie mises à la disposition de l'internaute. Ce sont par exemple, des applications sur Google Maps ou Bing Maps (Microsoft), OpenStreetMap qui permettent d'afficher de nombreuses données de base différentes (routes, images satellites, topographie, adresses). Sur ces applications l'utilisateur dispose de nombreux outils de manipulation des données spatiales pour soit créer de la nouvelle donnée géographique (création d'objet) soit modifier celle existante (mise à jour, enrichissement, qualification) soit encore personnaliser ces applications (mise en œuvre de fonctionnalités, échelle, type de données) (Roche et al., 2013).

Puis, l'utilisation de solutions mobiles (Smartphones munis avec des fonctionnalités de localisation de type GPS, Wifi, 3G) a permis aux utilisateurs de géolocaliser leur environnement et leurs activités en étant motivés par la volonté de mieux appréhender l'espace et les lieux, parfois en temps réel. Par exemple, les utilisateurs de GPS autonomes se servent de la technologie Map-share de Tom-Tom pour effectuer des modifications en temps réel (routes en travaux, nouveau commerce, nouvelles rues).

Enfin, la production explicite de l'information géographique apparait dans le cadre des initiatives professionnelles à travers des groupes de personnes qui sont invités à enrichir, mettre à jour et corriger les bases des données existantes institutionnels(Ripart de l'IGN) (Jolivet et Olteanu-Raimond, 2017) ou privés (Map Maker de Google, Mapinsight de Tele-Atlas ou Map Reporter de Navteq) (Hayat, 2019). Cela s'appuie sur le principe de crowdsourcing désignant de la production par la foule en profitant de la créativité et du savoir-faire des utilisateurs (Pisani et Piotet, 2008).

Par contre, dans la cadre d'une production implicite, on peut se servir des outils de géolocalisation et de géocodage pour retrouver la localisation des contenus (photo, vidéo, article, lien) publiés par l'utilisateur pour créer de l'information géographique. L'utilisation de mashups cartographiques pour traiter tout type d'information (incendies, tremblements de terre, etc.) en temps de crise est très récente et devient de plus en plus systématique. L'information est géocodée, puis intégrée sur la carte en ligne afin de suivre les évènements presque en temps réel (Roche et al., 2013).

En dernier, l'acquisition de l'information géographique peut être issue de l'importation et l'intégration de données géographiques issues des bases de données institutionnelles dans les projets volontaires comme OpenStreetMap.

A l'instar de la cartographie professionnelle, cette information géographique générée par les utilisateurs constitue, sans aucun doute, une nouvelle source de savoirs locaux, assez informelle certes, mais d'une telle richesse potentielle qu'elle représente aujourd'hui un complément pertinent aux données institutionnelles (Seeger, 2008; Heipke, 2010).

2.2.3 Comparaison de l'acquisition traditionnelle et volontaire

La comparaison de deux types d'acquisition de données s'opère tout d'abord au niveau de la démarche de collecte. Un jeu de données collecté de manière traditionnelle obéit à un processus ou un protocole dont la technique de collecte est déjà spécifiée et doit demeurer uniforme tout au long du processus d'acquisition, tandis que pour un jeu de données volontaires, ce mécanisme peut ne pas être respecté par les contributeurs. De ce fait, l'amplitude des erreurs semble être homogène et cohérente pour un jeu collecté de manière traditionnelle alors que pour les données volontaires, du fait de la variété des appareils de mesures, la qualité est variable d'une donnée à une autre, rendant ainsi les données volontaires d'autant plus hétérogènes. De la même manière, la collecte des données traditionnelles a vocation à respecter une forme d'homogénéité à la fois spatiale et temporelle dans la couverture d'un territoire, et est soumise à une planification des zones d'acquisition alors que les données volontaires sont collectées de manière relativement indépendante entre les contributeurs, ce qui peut engendrer une disparité dans la complétude spatiale de ces données.

Du point de vue de la précision des appareils de collecte, il parait en général que les données traditionnelles semblent plus précises que les données volontaires. En effet la précision des données conventionnelles varie entre des ordres de grandeur de 10 cm et de 1 m (e.g. la BDTOPO de l'IGN ayant une résolution de 1 m (Girres et Touya, 2010)). Les données volontaires sont en général numérisées à partir d'images de capteurs aériens (sur les plateformes web cartographiques) avec une erreur de position qui est souvent supérieure à 3 m (Barrette et al., 2000) ou à partir des récepteurs GNSS grand public ayant de précisions de positionnement variant de 5 à 15 m (Unger et al., 2013).

Par ailleurs, pour l'acquisition traditionnelle, les erreurs engendrées par les appareils de mesure lors de l'acquisition traditionnelle sont connues et donc l'incertitude engendrée par ces appareils semble être connue et quantifiable. Par conséquent, la mesure de la qualité des données apparait aussi quantifiable et objective. Par contre, pour la donnée volontaire, il arrive qu'on ignore le processus de production et de ce fait, une mesure de qualité semble plus difficile à établir. De plus, les opérateurs de la donnée traditionnelle suivent non seulement un protocole de mesure mais disposent également des compétences techniques. Ce qui renforce la crédibilité de la production des données contrairement aux contributeurs de projets volontaires où des nombreux des contributeurs sont qualifiés de non experts.

En revanche, la collecte des données traditionnelles nécessite beaucoup de temps. Il faut en général une phase de planification de la mission de terrain. On doit se munir des points d'appui sur le terrain ; il faut parfois les retrouver physiquement (balises géodésiques), puis saisir des points de liaison. On prévoit aussi du temps pour le calage et calibrage des appareils tout en ayant déjà choisi des conditions météorologiques favorables à la mission de terrain. Toutes ces contraintes et ces temps à ménager ne s'imposent pas pour la collecte des données volontaires. Leur saisie peut être plus facile et plus rapide que celle des

données traditionnelles.

En outre, les techniques d'acquisition des données traditionnelles demeurent assez coûteuses. Ce qui limitera la périodicité de la collecte de données. Il est également couteux de pouvoir identifier les changements qui ont eu lieu sur le terrain et qui doivent être intégrer dans la base de données. C'est pourquoi la mise à jour de ces données se fait plus lentement. Quant aux données volontaires, leur mise à jour est accrue et est presque instantanée (dès qu'il y a un nombre suffisant de contributeurs contribue sur une zone). Cela fait, en partie, l'intérêt des données volontaires. En effet, une donnée volontaire peut disposer plusieurs versions dans le temps issues de plusieurs contributeurs. Il existe également une communauté de contributeurs actifs dont le rôle est de renforcer, confirmer ou encore réfuter des contributions faites sur un lieu donné. Pour certains projets (comme OpenStreetMap) tout contributeur se voit attribuer une notation selon que ses contributions antérieures sont modifiées ou approuvés par ses pairs lui procurant une certaine hiérarchie et donc plus de confiance pour ses contributions. Ce qui en principe pour le projet OSM, garantit une certaine assurance de qualité et améliore les erreurs sur la donnée volontaire.

Enfin la différence entre les données produites par les volontaires et les données traditionnelles réside dans l'expertise du collecteur de l'information géographique (Coleman et al., 2010) :

- Dans le premier cas, on a un contributeur VGI qui a bonne connaissance locale du lieu à cartographier, capable de fournir beaucoup d'attributs sur le lieu ou sur les objets à cartographier et de renseigner leur historique de changement, qualifié pour une bonne utilisation des appareils GPS et des logiciels de cartographie du Web. Ce qui fournit des données rapidement capturées, en grande quantité et enrichies attributairement. Par contre, certains contributeurs ne connaissent pas les procédures du positionnement et les limites du GPS. Ce qui est susceptible d'engendrer des erreurs de position.

- Dans le second cas, on a un technicien qualifié maitrisant le protocole d'acquisition d'une agence de cartographie, ayant une bonne connaissance des limites d'utilisation de tout appareil d'acquisition, capable d'effectuer le relevé d'une donnée géographique précise. Il dispose également d'une bonne connaissance de la classification et de la délimitation des objets géographiques dans une base de données.

On a donc une donnée géographique volontaire abondante et facile à capturer avec plus d'attributs mais moins précise géographiquement et sémantiquement, dont on ignore le protocole de collecte suivi et dont l'objectif principal est de capturer, *ce qui est nouveau*, *ce qui a changé* et *ce qui se passe en temps réel*. Dans une quête de profiter de la richesse des données géographiques volontaires, il est important d'étudier les différents projets VGI existants pour dégager leurs potentialités pour des utilisations précises.

2.3 Initiatives des projets collaboratifs pour la collecte de données géographiques

2.3.1 Présentation des différents projets VGI

Un projet VGI est maintenu par une communauté de contributeurs ayant le but soit de cartographier un espace, soit de localiser des événements qui se passent dans cet espace. Ces contributeurs veillent à la maintenance, à la mise à jour et à l'amélioration des bases de données du projet. L'information géographique volontaire représente une excellente occasion d'améliorer au moins le processus de détection des changements et, à l'avenir, de créer des banques de données plus actualisées avec un étiquetage et des attributs plus riches. Coleman et al. (2010) recensent quatre contextes relatifs à quatre objectifs différents dans lesquels les individus contribuent volontairement à des informations géographiques dressant quatre catégories de projets VGI à savoir :

- *Cartographie et navigation* où l'objectif peut être une contribution à une série de cartes publiques (par exemple, USGS National Map Corps [4] est un projet de cartographie en ligne basé sur le processus de la production par la foule où des volontaires éditent des structures dans les États d'Amérique). Aussi les sociétés de GPS et de cartographie utilisent désormais les commentaires de ses utilisateurs pour améliorer tous les aspects liés aux informations routières, aux alertes de circulation, et aux avertissements de limitation de vitesse...etc. Ces commentaires constituent des contenus générés par les utilisateurs pour les fournisseurs des cartes de navigation. Dans la technologie Map-Share de Tom-Tom, il est possible d'apporter certaines modifications instantanément à l'aide du GPS(routes en travaux, nouveau commerce, nouvelles rues). Depuis son lancement en 2007 jusqu'à environ 2011, la technologie a comptabilisé plus de cinq millions de corrections de la part des utilisateurs (Mericskay et Roche, 2011).

  Un exemple d'utilisation d'un projet VGI porte sur la définition d'un itinéraire optimal proposé par un service de navigation. John et al. (2017) ont utilisé des traces GPS issues du projet collaboratif OpenStreetMap pour étudier l'inclinaison des surfaces afin de détecter dans un réseau routier les segments de forte pente et ainsi proposer des itinéraires qui conviennent au mieux à la navigation à faible consommation d'énergie et aux personnes à mobilité réduite. Malgré la faible précision absolue de ces données (les traces GPS), la précision relative des traces semble être suffisante pour calculer les valeurs d'inclinaison avec une précision raisonnable. Les auteurs estiment que face aux technologies d'acquisition traditionnelle des modèles numériques d'élévation (LiDAR aérien, photogrammétrie et levés terrestres) très coûteuses, cette approche issue du projet VGI constitue une alternative assez légitime tout en affirmant lors de la validation des résultats que la précision des valeurs d'inclinaison calculées à partir des traces GPS est légèrement meilleure que celle des valeurs d'inclinaison dérivées du MNE SRTM-1 (modèle numérique d'élévation sous licence ouverte). Cela montre qu'un projet VGI a bien permis d'améliorer le service rendu par une carte de navigation pour laquelle la précision relative des données géographiques est plus importante que la précision absolue.

4. https ://www.usgs.gov/core-science-systems/ngp/tnm-corps

- *Réseaux sociaux* où la contribution peut être apportée à un site comme (par exemple) le Christmas Bird Count, OpenStreetMap, Platial.com, Wayfaring.com, etc.

  Il existe des sites qui collectent de grandes quantités de données géographiques volontaires dont le but est d'enrichir les données géographiques d'une base de données régie par une licence libre d'utilisation et de partage des données. Il s'agit de sites comme *GPSPassion* ou *POIfriend* qui centralisent des milliers de bases de données de points d'intérêt (POI) sur des thématiques diverses (restaurants, radars, boîtes à lettres, parkings, etc.). Ces balises d'information complètent des données de base contenant les rues et les adresses. En complément des POI, des sites comme *Utagawa*, *Trailguru* ou *TraceGps* offrent des itinéraires pour des randonnées pédestres, cyclistes ou équestres sous la forme de tracés (GPS) consultables sur le Web et utilisables au sein des dispositifs mobiles (PDA, smartphones et GPS autonomes).

- *Civique/Gouvernemental* où le contributeur soutient une action en tant que citoyen concerné d'une ville donnée comme par exemple la contribution à un PPGIS (un Système d'Information Géographique sur la Participation du Public), ou membre d'un groupe de défense de l'environnement ou des droits des animaux.

  Par exemple, les projets de science citoyenne relatifs à la qualité de l'environnement et aux mesures de pollutions sont présents dans de nombreuses grandes villes du monde. Ces projets permettent d'une part de recueillir un grand nombre de données en multipliant les sources et d'autre part, de sensibiliser les citoyens aux problématiques environnementales urbaines. En France, le projet Montre Verte 10 illustre bien les potentialités de ce type d'initiatives basées sur le concept de citoyen capteur (Plantin, 2009). Ce programme cherche à mesurer de manière collective la pollution en milieu urbain (taux d'ozone et niveau de bruit). Cette application permet à l'utilisateur de visualiser en temps réel, sur une carte en ligne, les données géolocalisées des niveaux de pollution à partir des parcours individuels des porteurs de la montre (Mericskay et Roche, 2011).

  Le projet Noise Tube propose à ses utilisateurs une application mobile pour capturer l'ambiance sonore environnante à l'aide du microphone intégré, et lorsque le téléphone le permet, de géolocaliser les mesures. Les visualisations inédites qu'offre ce type d'applications permettent entre autres de mieux caractériser les zones et de proposer des solutions pour lutter contre la pollution sonore (Maisonneuve et al., 2009).

  Les travaux de See et al. (2016) donnent un aperçu du rôle du citoyen dans l'approvisionnement en masse de l'information géographique, en passant en revue les expressions utilisées pour décrire la perception des données géographiques par les citoyens ainsi qu'une exploration sur leur utilisation au fil du temps, avant de les classer par catégories et de mettre en évidence les questions clés dans l'état actuel de la science citoyenne. À l'occasion, les auteurs explorent différents projets VGI.
- *Rapports d'urgence* : lorsque la contribution soutient le signalement de la présence et de l'étendue d'accidents majeurs, d'incidents et de catastrophes naturelles ou d'origine humaine (incendies, inondations...) et autres événements de grande ampleur d'origine humaine (protestations).

Face aux grandes catastrophes naturelles récentes (cyclones, séismes, incendies, inondations), de nombreuses applications cartographiques Web destinées à la remontée d'informations pour aider à la gestion de crise ont fait leur apparition (Roche et al., 2013). C'est le cas de la plate-forme Ushahidi mise en ligne trois jours après le séisme en Haïti. Elle permet d'une part, de diffuser des informations sur le déroulement et l'évolution de la crise (demandes de secours, urgences logistiques, menaces potentielles, nouvelles individuelles, etc.), et d'autre part, de fournir un ensemble d'outils pour témoigner de la situation vécue de manière géographique et solliciter des interventions (SMS, courriel, formulaires en ligne, application pour téléphone mobile) en mettant rapidement en place un partenariat avec le fournisseur de téléphonie mobile et diverses organisations (entreprises, services de secours, ONG) (Mericskay et Roche, 2011). À la suite de cette initiative, le projet *Humanitarian OpenStreetMap Team* se forme pour répondre aux besoins des humanitaires et assister la cartographie des pays en voie de développement.

Dans tous les cas, les contributions faites dans chacune des catégories diffèrent selon l'objectif de contribution. Dans le cadre de cette thèse, on s'intéresse à la catégorie Réseaux sociaux, mais surtout du point de vue des données et pas des contributeurs, le cas du projet OpenStreetMap pour aboutir à une analyse des usages supplémentaires possibles du projet OSM. Dans un contexte d'absence de données de référence, OpenStreetMap pourrait être un cas d'application à partir duquel on constitue un référentiel topographique. Par la suite, on se focalise sur l'étude du projet OSM afin de connaitre les profils de ses contributeurs, leurs motivations pour ensuite dégager des utilisations possibles des données OSM.

2.3.2 Présentation du projet collaboratif OSM

Le projet OpenStreetMap est un des projets VGI les plus significatifs entretenu par une communauté active et structurée autour de l'usage du géoweb 2.0. Ce projet collaboratif de création d'une base de données libre et ouverte entre autre sur des rues, des routes, des points d'intérêts, des bâtiments du monde s'appuie essentiellement sur des contributions volontaires d'amateurs.

OpenStreetMap a été créé à la base par Steve Coast pour ne pas subir les restrictions interdisant la réutilisation, la reproduction ou la modification des données géographiques provenant des agences nationales de cartographie (Ordnance Survey pour la Grande Bretagne). Il a pensé à un monde de publication Libre des données géographiques où l'internaute est à la fois contributeur, utilisateur et diffuseur de l'information géographique sous une Licence Libre des contributeurs (ODBL).

De nos jours, le projet OpenStreetMap regroupe désormais une communauté de plus de 1,4 millions de contributeurs distincts (sans tous être des contributeurs actifs) à travers le monde à la date du 2020-02-07. Avec 3,5 millions de changements cartographiques par jour (2020-02-07), OSM demeure le projet le plus populaire et le plus actif du monde Libre [5]. Tout de même, on doit prêter attention à ces chiffres car les travaux de Neis et al. (2013) ont montré que les efforts de collecte de données ainsi que le nombre de contributeurs diffèrent significativement dans les régions du monde. Les auteurs ont pu observés des similitudes entre la densité de population des zones urbaines et la proportion des groupes

5. https ://www.openstreetmap.org/stats/data-stats.html

de contributeurs et le nombre de changements qu'ils ont apportés au projet OSM. De plus des facteurs socio-économiques, tels que le revenu, semblent avoir un impact sur le nombre de contributeurs actifs et les données fournies dans les zones analysées.

Par définition, une contribution sur OSM géolocalise un objet en fournissant deux types d'information : l'une géométrique (nœud, ligne, relation), l'autre attributaire (tag). Un nœud (node) est un point ayant une latitude et une longitude. Une ligne (way) est constituée d'un ensemble de points reliés les uns aux autres. Elle forme une surface (area) lorsque les points de départ et d'arrivée se rejoignent. Les relations sont utilisées pour regrouper des objets liés géographiquement. Enfin, un attribut (tag) est une paire clé/valeur utilisée pour décrire un objet. Par exemple, le tag *amenity=school* décrit une entité géographique dont la nature est une école primaire.

Le succès d'OpenStreetMap tant au niveau de la couverture géographique que de la qualité des données provient bien des possibilités offertes par la cartographie 2.0 (Girres et Touya, 2010; Haklay et al., 2010). Puis l'apparition des appareils de GPS moins chers, de précision assez bonne a favorisé a favorisé la création par les utilisateurs du Web de la donnée géographique volontaire dans OpenStreetMap qui se charge automatiquement à la mise en ligne de la nouvelle donnée géographique. Enfin OSM s'est popularisé grâce à l'apparition d'une communauté de citoyens capteurs (du débutant ou nouveau venu au géographe de niveau expert ou au développeur de logiciels) voulant non seulement cartographier tout changement dans leur entourage mais voulant participer à tout projet de gestion territoriale (Mooney et al., 2017). Derrière la plateforme OSM, il existe une communauté de developpeurs motivés pour lesquels la plateforme propose beaucoup d'outils d'édition, d'analyses et d'evaluationd ela qualité de données.

Les contributeurs d'OSM sont variés en terme de profils d'age et de nombre de contributions. (Coleman et al., 2010) a classifié les contributeurs d'OSM selon leur expertise, et l'intérêt qu'ils portent au domaine de Information Géographique :

- *Néophyte* : personne n'ayant pas d'antécédents formels dans un domaine, mais ayant de l'intérêt, le temps et la volonté d'offrir une opinion sur un sujet. Ce type de contributeur a identifié les lacunes dans la couverture cartographique, connaît bien le lieu et a obtenu l'équipement GPS nécessaire. Il est intéressé à apporter une première contribution ;

- *Amateur intéressé* : personne qui a *découvert* son intérêt pour un sujet, a commencé à lire la documentation de base, a consulté d'autres collègues et experts sur des questions spécifiques, expérimente son application et acquiert de l'expérience dans l'appréciation du sujet. Ce type de contributeur dispose de l'équipement de collecte, connaît les logiciels et les processus d'édition de données. C'est un contributeur régulier de données cartographiques éditées et il peut évaluer d'autres contributions ;

- *Amateur expert* : quelqu'un qui peut en savoir beaucoup sur un sujet, qui le pratique avec passion à l'occasion, mais qui ne s'en sert pas encore pour gagner sa vie. Ce contributeur maitrise l'équipement de collecte. Il évalue et édite régulièrement les contributions des autres sur ce sujet. Il participe à l'élaboration des spécifications et à la prise de décision ;

- *Expert professionnel* : quelqu'un qui a étudié et pratiqué un sujet, qui s'appuie sur ces connaissances pour gagner sa vie, et qui peut être poursuivi si ses produits, ses opinions et/ou ses recommandations s'avèrent inadéquats, incorrects ou diffamatoires. C'est un professionnel de la cartographie ou des services de géolocalisation.

D'autres chercheurs ont classifié les contributeurs selon le degré d'importance ou de la valeur ajoutée de leur contribution. Ils distinguent de contributeur selon le degré d'intervention humaine (opération de modification pourrait-elle être répertoriée uniquement par un humain ?), selon la fréquence, le type et le degré des opérations d'édition d'un contributeur, selon la réputation du contributeur (si la réputation de fiabilité d'une personne en termes de contributions et de révisions passées influence ou non la "durée de vie" des contributions ultérieures) (Priedhorsky et al., 2007).

Un enquête s'intéressant à la communauté OSM France a été réalisée dans le cadre du projet scientifique *ECCE Carto*[6] entre 2014 à 2017, une enquête a étudié la communauté OSM en France. L'enquête menée au sein du laboratoire PASSAGES (UMR 5319) du CNRS révèle un portrait du contributeur type. Le contributeur-type d'une communauté OSM (en France) est masculin et trentenaire, doté d'un niveau de diplôme élevé, plutôt cadre et *évoluant entre informatique et territoires*, avec un *intérêt marqué et renforcé pour les cartes*. Il est volontiers, collaboratif et ouvert *sur le Monde* . Plus précisément, pour 298 réponses obtenues, 88 % des répondants sont des hommes (chez les plus jeunes contributeurs la part homme/femme est un peu plus équilibrée, mais les contributeurs de 27 à 39 ans sont presque tous des hommes), la moyenne d'âge du contributeur est de 38 ans et 58 % des répondants ont un niveau master ou supérieur ; 31 % sont ingénieurs dans le secteur privé et 41 % dans la fonction publique. source wikipedia.

Par ailleurs, d'autres chercheurs ont aussi constaté cette inégale participation des contributeurs dans OSM. Neis et Zielstra (2014) ont analysé des études sur l'inégalité de la participation pour le projet OSM et ont constaté que 10% des personnes inscrites en 2008 ont contribué activement tandis qu'une étude de 2010 a montré que seulement 3,5 % des volontaires représentaient 98 % du contenu (Neis et Zipf, 2012).

Partant de ce constat, d'un contraste entre types de contributeurs, et entre proportions de participation, des chercheurs tentent de dégager des pistes de réflexion pour augmenter le nombre actif des participants d'OSM en analysant les motivations des contributeurs. Les travaux de Fritz et al. (2017) affirment que la compréhension de ces motivations peut fournir des stratégies pour transformer les cartographes occasionnels en cartographes plus sérieux, et aussi des moyens qui peuvent aider à renforcer la confiance des contributeurs occasionnels et souligner l'importance et les points forts des connaissances locales.

L'étude des motivations des contributeurs est évoquée d'abord par Coleman et al. (2010), puis par Budhathoki et Haythornthwaite (2013) et enfin dans Fritz et al. (2017) où une synthèse de motivations nous est livrée.

L'analyse de Coleman et al. (2010) sur les motivations du contributeur est basée sur la nature de l'intention du contributeur à savoir *altruisme*, *intérêt professionnel ou personnel*, *stimulation intellectuelle*, *protection ou valorisation d'un investissement personnel*, *récompense sociale*, *amélioration de la réputation personnelle*, *participation offrant*

6. https ://www.data.gouv.fr/fr/organizations/projet-de-recherche-ecce-carto/

un moyen d'expression créative et indépendante, et *fierté d'appartenance"* (incluant la connaissance locale). Par exemple, la connaissance locale est bien répandue dans OSM où les cartographes cartographient ou mettent à jour plus fréquemment leurs zones locales que des zones plus éloignées, à moins qu'ils ne soient motivés par la carto-partie (mapping party) ou des causes humanitaires.

Dans Budhathoki et Haythornthwaite (2013), les auteurs distinguent les motivations intrinsèques qui viennent directement de l'individu, et les motivations extrinsèques, qui viennent de l'extérieur telles que les incitations financières ou l'acquisition d'une réputation positive basée sur la qualité de ses contributions ou de celles de ses pairs. A l'issue d'un enquête mené sur les volontaires d'OSM, les auteurs parviennent à dresser deux types de volontaires, à savoir les cartographes sérieux et les cartographes occasionnels, en fonction du nombre de contributions, de la durée des contributions ou de la fréquence des contributions. Les résultats de l'enquête menée auprès des 444 volontaires d'OSM ont montré que deux facteurs extrinsèques, tels que la communauté et l'objectif du projet, et les facteurs intrinsèques d'ethos et d'altruisme étaient les plus importants (Fritz et al., 2017).

Une dernière forme de motivation, mais moins fréquente pourrait être animée par une intention d'ordre vandale (espérant générer le scepticisme ou la confusion en remplaçant les entrées légitimes par des contenus absurdes ou ouvertement offensants), d'ordre de croyance ou de conviction ou encore malveillante/criminelle (personnes ayant une intention malveillante (et éventuellement criminelle) dans l'espoir d'un gain personnel) conduisent à des contributions de mauvaise qualité (Truong et al., 2019b; Coleman et al., 2010).

Dans cette thèse on ne traite pas les contributions faites dans le cadre du vandalisme non plus des profils d'utilisateurs et leurs motivations.

2.3.3 Analyse d'utilisations possibles des données OSM

Le projet OpenStreetMap possède une communauté très active et nombreuse avec une base de données parfois plus complète que les bases de données d'autorité (Girres et Touya (2010). Ce double avantage séduit les agences nationales car non seulement les données d'OSM peuvent servir de sources pour la mise à jour des données gouvernementales mais les contributeurs peuvent également être amenés à connaître davantage les données d'agences, et éventuellement à les utiliser et à les modifier à travers des plateformes dédiées et contrôlées par les agences nationales dans le cadre des projets collaboratifs proposés par les agences. D'après Olteanu-Raimond et al. (2017b), Le Libre accès aux données d'autorité a permit d'accroître la facilité d'utilisation des données faisant autorité (les utilisateurs téléchargent et utilisent ces données à différentes fins) et d'augmenter la participation des citoyens et des partenaires à l'édition des données d'autorité, avec la possibilité d'ajouter de nouvelles informations, de donner un retour d'information et d'émettre des alertes sur les erreurs et les mises à jour. Les auteurs citent l'exemple des collectivités locales qui incitent les citoyens, à la fois en tant que capteurs et en tant que partenaires potentiels, à collecter des données à différentes fins (comme dans le domaine de l'urbanisme, afin de faire connaître les nouvelles réglementations) (Karimipour et Azari, 2015; Sedano, 2016)).

En outre, certaines agences nationales de cartographie européennes souscrivent à la politique du Libre accès (Brovelli et al., 2016). Ces agences utilisent une licence de données qui autorise à une tierce personne, d'utiliser les données d'autorité, puis se l'approprier

pour intégrer les données des agences dans OSM. Il s'agit d'un import souvent massif dans la base OSM. Si cela permet de rajouter d'un seul coup beaucoup de données dans OSM, cette manipulation menace le caractère volontaire des données OSM et elle diminue l'apport des contributeurs de la base. Et si la licence adoptée par certaines agences favorise l'amélioration des données OSM, en revanche rien ne garantit que cela se traduise *in fine* par une amélioration réciproque sur les données des agences.

Par ailleurs, Olteanu-Raimond et al. (2017a) rapporte qu'à travers une enquête réalisée sur 23 agences nationales de cartographie en Europe, 12 agences utilisent fortement l'information géographique volontaire et que plusieurs parmi celles qui ne l'exploitent pas sont prêtes à s'y mettre dans un futur proche. Ces utilisations s'opèrent dans 5 volets ou contextes : la détection des changements, les alertes de signalement, la collecte de nouveaux contenus, les noms de lieux vernaculaires et l'interprétation de photos. La plupart des agences ont utilisé les données volontaires pour détecter des changements et collecter des corrections d'erreurs tandis que 2 agences intègrent l'information géographique volontaire pour la collecte des nouveaux contenus géographiques. 9 agences ont prévu soit de commencer à collecter du VGI (six agences), soit d'étendre davantage ce qu'elles collectent actuellement (trois agences). Toutefois, cinq obstacles majeurs à l'utilisation du VGI ont été identifiés par les agences : la qualité et la validation des données (contrôle de qualité),les questions juridiques de droits des données et responsabilité, l'expertise et la motivation de la foule, la motivation à long terme et enfin les craintes en matière d'emploi.

Néanmoins, toutes ces initiatives nécessitant un cadre bien contrôlé (avoir des outils automatiques de corrections d'erreurs), la plupart des agences utilisent leurs propres outils de collecte du VGI(allant des outils pour des projets de recherche à des outils éprouvés et pleinement opérationnels) à la place des outils VGI habituels dans la communauté comme celui par lequel on collecte les données OSM (plateforme d'acquisition exemple). Cela suppose d'avoir un groupe de citoyens spécifiques formés à ces outils.

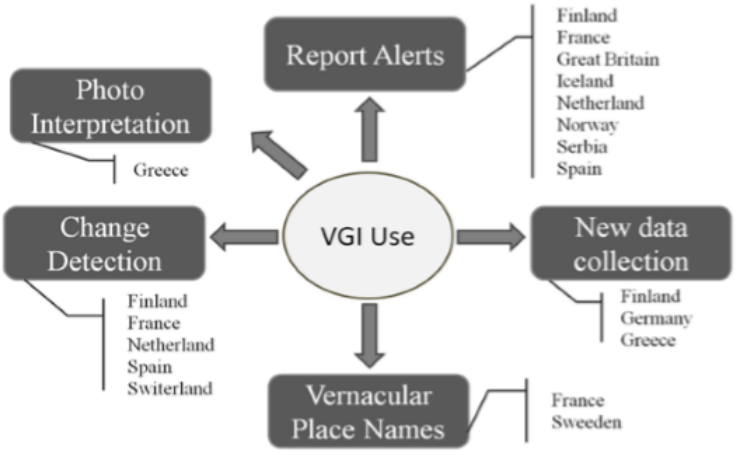


FIGURE 2.1 – Utilisation des VGI dans les agences nationales européennes de cartographie ; Source : (Olteanu-Raimond et al., 2017b)

Quelques exemples d'utilisation et d'intégration du VGI dans les données d'autorité

sont donnés selon leur contexte dans les travaux de Olteanu-Raimond et al. (2017a) :

- Détection de changement : Aux Pays-Bas, "Kadaster" a entrepris un test pilote pour la détection des changements concernant les routes et les bâtiments dans son ensemble de données 1 : 10k (TOP10NL) à partir de sources externes (par exemple les ensembles de données des administrations locales et OSM). Ce test a montré que les changements ont été détecté plus rapidement qu'avec les méthodes utilisées traditionnellement à "Kadaster".

- Correction d'erreurs : l'IGN France a développé un système d'alerte de signalement appelé Ripart, qui est en service depuis 2008 (Viglino, 2009). Il s'agit d'un système d'approvisionnement communautaire dans lequel des partenaires professionnels, tels que les pompiers, proposent des mises à jour des données de l'IGN en remplissant un formulaire et en fournissant des informations de localisation à l'aide de traces GPS, de photographies ou de dessins sur une carte. Le processus de notification de Ripart fournit des croquis se chevauchant des changements proposés sur une carte de base inerte. Ce service peut également être utilisé par les partenaires pour leurs propres besoins (mise à jour de données métiers, non destinées à l'IGN).

- Noms de lieux vernaculaires : l'agence nationale de cartographie et l'agence de cadastre et d'enregistrement des terres de la Suède, ont mis en place une application pour smartphone, appelée "Platsnamna", pour collecter les noms de lieux vernaculaires (noms populaires des lieux parfois différents des noms usuels utilisés par les agences), qui a été testée dans des zones urbaines, avec le soutien des professionnels et du public.

- Collection des nouvelles données : l'idée consiste à saisir des données sur des nouveaux objets géographiques ou attributs étant auparavant exclus du cadre de mission (d'ordre économique ou politique) des agences de cartographie. L'Agence fédérale allemande de cartographie et de géodésie utilise le VGI pour collecter de nouveaux contenus, caractéristiques ou attributs, tels que des bâtiments, des structures (par exemple, ponts, pylônes, silos, barrages), des routes, des pistes et des sentiers, des caractéristiques hydrographiques (par exemple, rivières, lacs, canaux), des noms, des points d'intérêt (par exemple, attractions touristiques, distributeurs de billets). Les nouvelles données sont collectées par des sources secondaires telles qu'OSM.

Toutefois, l'implication des agences européennes se voit limitée dans l'intégration de l'information géographique volontaire par des préoccupations liées à des questions de qualité et de validation des données, des questions juridiques, des questions liées à la nature et à la motivation de la foule, des questions de durabilité et des craintes liées à l'emploi (Olteanu-Raimond et al., 2017b). Ces questions constituent des challenges futurs dans le domaine de la recherche portant sur l'étude des potentialités des projets volontaires, notamment OpenStreetMap.

2.4 Contexte et problématique de la thèse

2.4.1 Contexte de la thèse

Dans le contexte de Djibouti, on assiste à une production et à une utilisation accrues et diversifiées des données géographiques par les différents acteurs/services qui souhaitent en tirer de l'information qui pourrait les aider à prendre une décision. Cependant ces productions et utilisations ne sont régies par aucune norme concernant leur diffusion ou la communication des métadonnées sur leur élaboration. Il est donc impossible de se prononcer quant à la qualité des données issues de ces productions. Il en est de même pour la fiabilité de toutes les études géographiques ou cartographiques qui s'y rattachent.

Il est ainsi fréquent qu'un département se serve des données géographiques issues d'une étude faite par un autre département ou service sans se soucier des conditions dans lesquelles les mesures ont été faites ou prises, et produise une nouvelle information ou donnée qu'il va à son tour passer à un autre acteur national. Ainsi, des erreurs s'accumulent, les données géographiques perdent toute leur fiabilité et donc de leur qualité. On ne peut en aucun cas se prononcer sur la fiabilité des cartes et modèles géographiques qui en résultent.

Cela est dû à l'absence d'une instance nationale qui s'occuperait de la création et de la gestion d'un référentiel cartographique qui servirait de base (topographique et thématique) pour toutes ces études. En temps normal, la plupart des pays évaluent et enrichissent leurs données géographiques nationales à travers des agences nationales dédiées à la collecte, l'organisation, la production et la mise à jour de ces données en élaborant des référentiels cartographiques. Ces référentiels cartographiques contiennent un ensemble de jeux de données validé par une instance officielle comme étant un produit géographique de référence sur lesquels peuvent s'appuyer de nouvelles constructions géométriques de précision moindre ou égale. Or, Djibouti fait partie des pays n'ayant pas encore mis en place un référentiel géographique national. La mise en place d'un référentiel national admet plusieurs avantages.

Tout d'abord, la disposition des données de référence relève d'une question de la souveraineté du pays, puis d'une question légale et d'opposabilité. En effet, une base de données de référence peut servir de document de référence pour régler des problèmes de chevauchement de parcelles et de ce fait peut permettre de faire valoir un droit à un dédommagement quand des installations ou infrastructures percutent des propriétés foncières. Actuellement, à Djibouti, bien qu'il existe une base de données cadastrale, cette dernière n'est pas en harmonie ou en accord avec les données des autres services publiques ou concessionnaires. Cette base ne peut pas permettre de résoudre un litige car il n'existe pas un document commun entre le cadastre et les autres services publics.

Cela nous amène à évoquer la question de l'interopérabilité. Deux bases de données seront dites interopérables si, grâce à une ou plusieurs norme(s) externe(s) qu'elles respectent, elles en viennent entre autres à pouvoir être compatibles. L'interopérabilité est générale et ne concerne pas à priori des données ou systèmes de données. Elle existe au travers de normes et formats respectés par tout élément ou système qui souhaite intégrer un plexus interopérable - le réseau des éléments (base de données) qui communiquent entre eux de façon fluide et normée. Elle doit résulter d'un accord explicite entre les différents acteurs ou services producteurs/utilisateurs des donnés géographiques. Cela exige de partir d'un référentiel unique et commun pour produire des produits métiers dérivées et propres

aux besoins d'un service donné. De plus, mettre en place un mécanisme d'interopérabilité entre plusieurs produits cartographiques permettrait d'être capable d'avoir un même formalisme et donc d'harmoniser la qualité sur les données géographiques. Cela peut rendre possible d'avoir une qualité commune sur un produit cartographique global et diffusé au sein des services nationaux à Djibouti.

Pourtant, le besoin de données de référence couvrant l'ensemble du territoire se fait de plus en plus ressentir d'avantage. En effet, une étude faite dans le cadre d'un projet de développement [7] estime que la carte topographique à grande échelle existante fournissant les données de base est devenue vieille et ne permet pas de cerner l'étendue de l'agglomération qui s'agrandit d'année en année au fur et à mesure de l'accroissement démographique. A ce sujet, les auteurs du rapport, précisent que différents donateurs tels que l'Union Européenne, la Banque mondiale, l'Agence française de développement etc., mettent en œuvre différents projets d'infrastructures tels que l'aménagement des réseaux d'eau et d'assainissement etc., mais faute de carte topographique, chaque projet réalise indépendamment la reconnaissance du terrain et le levé topographique. Cela montre bien encore là l'intérêt de disposer d'un référentiel national qui permettrait d'économiser les efforts répétitifs des levés topographiques localisés sur des zones, au profit d'une mise à jour assez étendue sur le territoire national.

2.4.2 Données géographiques de référence

Après l'étude du contexte de Djibouti, on constate qu'il existe une confusion entre les appellations telles que *données de base* ou *données souveraines* ou encore *données de référence* ainsi que d'autres termes qui s'y rattachent.

En effet, on qualifie des données souveraines selon le rapport [8], des données servant de support direct aux décisions de la puissance politique. Les données géographiques souveraines fondent l'état de droit : le parcellaire, les servitudes, les règlements d'urbanisme et environnementaux. Elles permettent d'établir une décision publique opposable, liée à l'État de droit. Dès lors que l'État ou l'administration peut prendre des décisions pouvant s'imposer aux citoyens, la donnée géographique souveraine doit présenter des garanties d'indépendance et d'autorité. Si l'autorité publique ne doit dépendre de personne pour disposer de la donnée souveraine, elle doit, si ce n'est la produire en totalité, à minima maitriser techniquement et continument son processus de production, de l'acquisition de base jusqu'au stockage de la donnée, en passant par son élaboration et sa mise à disposition. De ce fait, la donnée souveraine ne doit pas être imposée par voie d'autorité, mais doit faire autorité par sa qualité et son caractère documenté.

Ainsi, selon ce même rapport [9], l'auteur estime que les données souveraines peuvent être issues des processus collaboratifs. Dans ce cas, les exigences de fiabilité qui s'attachent à ce type de données impliqueront parfois que les contributeurs soient sélectionnés et identifiés comme de *tiers de confiance* (par exemple : collectivités territoriales, services

7. rapport sur un *Projet de gestion de données topographiques numériques A Djibouti ville* rédigé par l'Agence Japonaise de Coopération Internationale (JICA), à la date Mars 2014

8. rapport sur les *données géographiques souveraines* rédigé par Madame Valéria FAURE-MUNTIAN, députée de la Loire, à la date juillet 2018

9. rapport sur les *données géographiques souveraines* rédigé par Madame Valéria FAURE-MUNTIAN, députée de la Loire, à la date juillet 2018

déconcentrés de l'État, délégataire d'une mission de service public). Les données collaboratives doivent aussi faire l'objet d'une analyse aux exigences de qualification modulables en fonction des sources mobilisées. Toutefois, plus la donnée sera mobilisée au service des missions régaliennes, qui constituent *le cœur* de la souveraineté, plus les possibilités offertes de myriadisation (crowdsourcing) seront restreintes.

Par contre les données de référence sont des données qui constituent une référence commune pour nommer ou identifier des produits, des services, des territoires ou des personnes et sont réutilisées fréquemment par des personnes publiques ou privées autres que l'administration qui la détient, et dont la réutilisation nécessite qu'elles soient mises à disposition avec un niveau élevé de qualité. Comme données de référence, en France, on cite le plan cadastral informatisé (PCI) géré par la Direction Générale des Finances Publiques, le Registre Parcellaire Graphique produites par l'Agence de Services et de Paiement (ASP) et enfin le Référentiel à Grande Échelle (RGE). Une donnée de référence [10] désigne la donnée publique dont le potentiel d'usage est élevé.

A Djibouti, certes des données souveraines existent, mais elles sont non unifiées. Le service de Cadastre dispose de ses propres données souveraines tout comme l'urbanisme et tous les autres services publics. Des échanges de données entre services publics se font de manière informelle et les métadonnées ne sont pas systématiquement transmises.

2.4.3 Problématique de la thèse

Depuis une dizaine d'années, la communauté de chercheurs s'intéresse à la qualité du VGI (Girres et Touya, 2010). Ceci trouve son importance du fait que la donnée VGI constituerait une très bonne source pour créer des données de référence vu sa richesse, sa disponibilité et son universalité. Cela pourrait servir de base à un processus de mise en place des référentiels nationaux pour les pays qui n'en disposent pas, tels que les pays du Sud.

Afin d'évaluer la qualité des données OpenStreetMap, les travaux de Siebritz (2014) ont évalué les données d'OSM en comparaison avec la base de données topographiques (en Afrique du Sud) afin mettre à jour cette base de données.

D'autres travaux montrent que les données OpenStreetMap peuvent servir de modèle pour constituer des données de base.

Ainsi, dans la cadre d'une collection de données participatives, il est proposé aux citoyens de participer à des campagnes de collecte de données. Par exemple, dans un cas mentionné par Kalantari et La (2015), des citoyens volontaires fournissent la cartographie des limites et des propriétés cadastrales. See et al. (2017) précisent que cela est particulièrement pertinent pour les pays en développement où les droits fonciers ne sont pas bien documentés. Cela est également pertinent dans les endroits où les levés sont très coûteux et longs et n'ont donc pas pu être effectués dans tous les domaines, ce qui entraîne une stagnation du marché immobilier.

10. rapport sur les *données géographiques souveraines* rédigé par Madame Valéria FAURE-MUNTIAN, députée de la Loire, à la date juillet 2018

Un autre exemple grec est présenté par Basiouka et Potsiou (2012), qui ont mené une expérience dans la partie rurale du village de Tsoukalades, sur l'île de Leucade, où quinze propriétaires fonciers volontaires ont utilisé un GPS portable pour délimiter leurs parcelles de terrain. Lorsque les résultats ont été comparés à une enquête officielle, les emplacements et les formes de toutes les parcelles se sont révélés corrects et la majorité des parcelles avaient des calculs de superficie qui étaient dans les limites de tolérance des spécifications fixées par le cadastre hellénique. Ainsi, la participation des citoyens offre un grand potentiel pour aider à recueillir ce type d'information cadre (See et al., 2017).

Dans une étude plus récente de Basiouka et al. (2015), des étudiants en arpentage ont été chargés d'évaluer la faisabilité de l'utilisation d'OSM pour la cartographie cadastrale à Athènes, en Grèce. Les résultats ont montré une bonne précision, un faible coût et une facilité d'utilisation pour les non-experts, ce qui indique que l'OSM est une solution possible pour l'externalisation des parcelles et des caractéristiques des terrains, en particulier si l'on adopte une solution hybride dans laquelle des experts en topographie sont utilisés pour la formation et l'assurance qualité (See et al., 2017).

Dans Olteanu-Raimond et al. (2017b), il est question d'étudier l'intégration du VGI dans des données d'autorité dont disposent les agences de cartographie. Les auteurs affirment que les autorités cartographiques grecques ont utilisé le VGI comme point de départ pour mettre à jour ou créer de nouveaux produits cartographiques. Les données provenant de la foule sont traitées comme une couche d'entrée initiale qui est comparée à des images de fond (satellites ou aériennes). Les ensembles des données volontaires sont corrigés, complétés et réaffectés à la nomenclature locale, puis suivent les mêmes processus pour les données collectées en interne. On a également les travaux de Ivanovic (2018) qui ont montré l'intérêt des traces GPS pour mettre à jour les données institutionnelles (BDTOPO de l'IGN).

Dans un contexte d'absence de données de référence ou d'autorité, ces exemple montrent bien qu'il est judicieux de considérer et d'étudier un cadre dans lequel la base de données OSM pourrait constituer une première ébauche de données en vue de mettre en place un référentiel cartographique basé sur OpenStreetMap.

Malgré ces efforts, la qualité des données VGI et notamment les données d'OSM demeure une question loin d'être résolue, elle anime beaucoup de débats de recherche surtout que la richesse des ces données ne cesse de s'accroître (Senaratne et al., 2017).

En effet, en temps normal, l'évaluation de la qualité de la base OSM se fait par comparaison avec une base dite de référence issue des données d'autorité en passant par une étape d'appariement qui consiste à retrouver des objets homologues dans les deux bases (base OSM et la base de référence). Cette comparaison s'établit à travers des indicateurs de qualité issus de la norme ISO qui entre autres, mesurent la précision spatiale exprimant l'écart de la base OSM par rapport à la base de référence en termes de position et de forme. Or du fait que les données d'autorité et OSM diffèrent de nature en matière de processus de production et de mise à jour, la base OSM peut être parfois nettement plus détaillée, plus précise et plus complète que les jeux de données faisant autorité comme déjà mentionné dans 2.2.3, violant ainsi l'hypothèse de base de l'utilisation des données d'autorité comme base de référence (Minghini et Frassinelli, 2019; Antoniou et Skopeliti, 2015). De plus, Minghini et Frassinelli (2019) affirment que dans des nombreuses pays

les données d'autorité soit soit manquantes, soit inadaptées à la comparaison(parce que leur échelle est trop grande et n'est pas comparable à celle d'OSM). Pour ces raisons, les auteurs soulignent l'apparition des méthodes d'évaluation intrinsèque (c'est-à-dire sans comparaison avec des ensembles de données tiers) pour la qualité d'OSM.

Ainsi des travaux de recherche ont fait appel à l'apprentissage automatique et ont utilisé des réseaux de neurones pour reconstruire au mieux un indicateur de qualité. C'est le cas des travaux de Xu et al. (2017) qui ont utilisé un auto-encodeur pour reconstruire une meilleure représentation de la précision spatiale mesurée entre des données OSM et une base de référence. Cette représentation est apprise à l'aide des neurones, en comparant l'empreinte des bâtiments par une image raster. Or, notre objectif ne consiste pas à retrouver une meilleure représentation de la précision spatiale mais plutôt de l'estimer à partir d'une évaluation de la qualité faite uniquement sur les données à un état donné (non pas l'historique sur les données). Cela permettra plutard d'estimer la précision spatiale sans à devoir recourir une base de référence. Ce qui nous amène à formuler la problématique centrale de notre thèse :

- Comment évaluer un jeu de données OpenStreetMap de manière intrinsèque et sans Référence en vue de l'adopter comme une référence géographique nationale ?

Dans la cadre de cette thèse, nous nous intéressons uniquement à évaluer la précision spatiale de la base OSM. Pour répondre à notre problématique, on propose un framework à partir duquel on infère la précision spatiale relative des données OSM à l'aide des méthodes de l'apprentissage automatique. Pour cela, on formule l'hypothèse qu'il existe un lien statistique entre la précision spatiale des objets géographiques et leur qualité de saisie.

Ainsi, on dresse les objectifs spécifiques ci-dessous pour aboutir à une évaluation de la précision spatiale d'un jeu de données OSM sans devoir le comparer par une base de référence :

- Définir des indicateurs extrinsèques en faisant apparier un jeu de données OSM avec une base de référence. Cette étape nécessite de définir un ensemble de mesures qui estiment les écarts de la base OSM à la base de la référence. Ces écarts constituent la précision spatiale relative des objets de la base OSM.

- Identifier et définir un ensemble d'indicateurs intrinsèques sur les objets OSM qui révèlent des erreurs géométriques de saisie et qui peuvent générer des connaissances sur la mauvaise ou bonne qualité de position et de forme de ces objets OSM.

- Identifier des méthodes d'apprentissages permettant de générer la précision spatiale des objets de la base d'OSM en s'appuyant sur les indicateurs intrinsèques.

- Étudier la Transférabilité du modèle d'apprentissage obtenu sur d'autres jeux de données issus d'OSM.
- Appliquer le modèle d'apprentissage final sur un jeu de données tiré de l'OSM de Djibouti.

0

Chapitre 3

État de l'art

Sommaire

La problématique de la qualité de l'information géographique est d'actualité et continue à susciter l'intérêt des chercheurs et praticiens en géomatique (à trouver des citations plus récentes).

Ce domaine de recherche a connu un développement important, grâce à l'émergence de l'informatique et à l'amélioration du réseau Internet. De même, l'accessibilité croissante des données et de dispositifs produisant de l'information géographique, et l'adoption de l'information géographique dans de nombreux domaines d'application ont contribué à ce développement (Devillers et al., 2007).

Les problèmes de la qualité touchent tous les domaines utilisant l'information géographique, que ce soit pour un ingénieur en environnement qui doit utiliser un modèle numérique de terrain afin de modéliser un bassin versant, pour un géomètre expert devant combiner

plusieurs données pour faire une mesure très précise d'un territoire donné, ou simplement pour une personne utilisant un site web qui permet de localiser une adresse sur une carte. On peut énumérer des exemples de l'influence possible de la mauvaise qualité des données dans une prise de décision dans le domaine du géomarketing (par exemple, une mauvaise implantation d'un centre commercial), de l'aide à la navigation (par exemple, une erreur de guidage d'un véhicule), de la gestion des risques (par exemple, un mauvais positionnement d'un ouvrage de protection). L'évaluation de la qualité sur la précision spatiale est primordiale dans tous ces domaines-là. Et son évaluation s'inscrit dans un concept de qualité couvrant l'ensemble du processus d'acquisition, gestion diffusion et utilisation de l'information géographique (Devillers et al., 2007). Cela nous oblige à reconsidérer de plus en plus sérieusement le problème de la qualité des données dont l'impact a une influence directe sur la fiabilité des analyses spatiales produites et des décisions qui en découlent.

Un autre besoin qui surgit est celui de la communication de la qualité. Il s'impose à tout producteur ou distributeur d'informer les utilisateurs sur la qualité des données spatiales de manière significative et compréhensible (comme n'importe quel autre produit ou service sur le marché), que ce soit dans un contexte professionnel ou pour une utilisation *grand public*.

Une connaissance fine et détaillée sur la qualité des données parait être un prérequis indispensable pour toute tentative d'évaluation de la qualité et de sa communication. Cette connaissance se situe en réalité à deux niveaux (le premier niveau est conceptuel tandis que le deuxième est plutôt méthodologique) :

- De quoi parle-t-on quand on parle de qualité des données géographiques ? Quels sont les différents aspects de la qualité ?

- Quelle démarche de mesure de qualité doit-t-on adopter pour évaluer la qualité des données géographiques de manière générale ?

La première question consiste à synthétiser l'ensemble des concepts ou termes qui se rapportent à la qualité de manière à identifier la nature de la qualité dont il est question face à chaque contexte, tout en respectant la chronologie d'apparition de ces concepts. Il est également question de dégager une définition de la qualité selon la nature des données et selon l'objectif recherché. Répondre à ces questions permet d'aider à la formulation des spécifications pour évaluer la qualité.

La seconde question tente de fournir une synthèse sur les démarches adoptées pour une évaluation de la qualité. On s'interroge de quelle manière les données produites se rapprochent ou s'écartent des spécifications initiales. Dans le cas ou les spécifications d'un jeu de données sont méconnues, il serait judicieux de procéder par comparaison avec avec un jeu de données dont la qualité est connue, ou encore de tenter de l'évaluer de manière intrinsèque. Aussi, la qualité peut porter sur une évaluation de la précision ou des incertitudes engendrées par l'utilisation d'une telle ou telle autre méthode ou enfin à une modélisation de la propagation d'erreurs ou d'incertitudes sur le produit final.

Cependant, ces démarches usuelles adoptées pour une évaluation de la qualité s'avère pas très adaptée pour le type de données dont on souhaite traiter dans cette thèse.

Ce type des données géographiques ont la particularité d'être capturées par des contributeurs bénévoles. Elles sont communément appelées les *données géographiques volontaires (Volunteered Goegraphic Information-VGI)*. De ce fait, nous nous livrons une étude méthodologique de synthèse sur leur qualification.

Dans ce chapitre, nous présentons tout d'abord le cadre théorique dans lequel nous tentons de résumer les terminologies et aspects autour de la qualité de l'information géographique suivi d'un cadre technique étudiant les différentes démarches d'évaluation de la qualité d'un jeu de données. Puis nous focalisons notre analyse sur la qualification des Enfin, nous détaillons tous les outils utilisées dans notre processus de qualification des données géographiques volontaires notamment les données de la base OpenStreetMap.

3.1 Cadre théorique sur la qualité des données géographiques

Quand on parle de qualité de données, nous avons plusieurs termes qui entrent en jeu, à savoir : exactitude, précision, assurance qualité, contrôle qualité, traitement qualité, amélioration de la qualité, paramètres qualité, erreur et incertitude. Ces termes seront énoncés de façon chronologique de façon à retracer l'historique du traitement de la qualité de données géographiques.

3.1.1 Concepts de la qualité des données géographiques au fil du temps

Durant ces trente dernières années, la perception de la notion de qualité de l'information géographique a bien évolué. Tout au début, dans le domaine de la cartographie, le terme le plus utilisé dans un contexte opérationnel était *Exactitude*. Par exemple, la norme nationale d'exactitude cartographique (National Map Accuracy Standards- NMAS (US Bureau of the Budget, 1947)) des États-Unis d'Amérique s'occupait exclusivement de l'exactitude spatiale, et uniquement pour des points bien définis, ces objets étant plus facilement testés de manière à vérifier si 90% des points testés rentrait dans un seuil de 0,5 ou 0,8mm selon l'échelle sur une carte papier. La qualité consistait alors à se conformer à ce seuil (Chrisman, 2005).

Plus tard, dans les années 80, ce test simple va être remplacer par l'adoption de concepts statistiques. On utilise l'écart-type et des tests de distribution. La différence est que la distribution des erreurs est traitée avec des outils statistiques. On détermine l'ecart-type de la distribution des erreurs sur les 90% de la population. Puis on calcule l'écart-type de la distribution normale théorique de ces erreurs en supposant que la distribution des erreurs suit une loi normale. Au lieu de spécifier que 90% des données doivent rester à l'intérieur d'un certain seuil, la norme d'exactitude circulaire cartographique (Circular Map Accuracy Standard), adoptée par le ACIC (l'agence cartographique militaire américaine de l'époque), prévoit de calculer la distance entre ces deux écart-type qui doit être inférieure à un seuil donné. En effet, la qualité selon le terme *exactitude* se mesure à travers le calcul des descripteurs statistiques sur un échantillon pris sur jeu de données. Les valeurs des descripteurs obtenus (écart-type) sont comparés à des seuils fixés au préalable (écart-type de la population sur la distribution théorique). Si leur distance (par rapport aux seuils fixés au préalable) dépasse à un certain seuil, la qualité du jeu de données est qualifiée de mauvaise.

Toutefois la qualité des données géographiques se concentrant initialement sur l'exactitude spatiale s'est développé dans de nombreuses directions. L'idée d'exactitude d'une carte doit être élargie afin d'englober l'ensemble du contenu informationnel. Cela signifie qu'il faut aborder *les attributs* d'une carte thématique, l'identité des objets d'un inventaire représenté sur un dessin d'ingénierie et de façon générale tout ce qu'un utilisateur pourra interroger d'où un nouveau terme de qualité, l'exactitude sémantique ou attributaire (Chrisman, 2005).

Par ailleurs, la gestion moderne de l'information a multiplié les opportunités pour les données de passer d'un utilisateur à un autre, pouvant alors éventuellement échapper aux limites des utilisations prévues. Un producteur veille à la qualité de sa production avec des tests et des seuils, mais il n'est nul part tenu responsable d'une mauvaise utilisation de la donnée géographique, et donc ne pourra garantir sa qualité, car le même produit pourrait être tout à fait utilisable pour un ensemble de clients, mais complètement inadéquat pour d'autres. Désormais, les utilisateurs ont également leur place dans le processus d'évaluation de la qualité. On assiste à une nouvelle notion de la qualité : *fitness for use*.

C'est pourquoi les producteurs ont dû communiquer plus tard les conditions d'utilisations et de création de l'information géographique dans un fichier appelé *métadonnées* afin que les utilisateurs puissent faire des jugements informés concernant l'adéquation à l'usage. Le terme *fitness for use* exprime la conformité de l'utilisation d'un jeu de données dans un contexte donné par rapport à l'utilisation prévue par le producteur et exprime donc ce qu'on appellera plus tard la qualité externe.

Au fil des années, les aspects de la qualité ont été enrichis en définissant la qualité autour de cinq composantes. La norme de transfert de données géospatiales (SDTS, Spatial Data Transfer Standard (National Institute of Standards and Technology, 1992)) fournit un rapport sur la qualité des données composé de cinq parties. La section *généalogie* décrit les sources et toutes les transformations faites pour obtenir le produit final. Les quatre autres sections décrivent des tests portant sur différents types d'exactitude. Le test *d'exactitude spatiale* utilise une méthode établie permettant de caractériser les différences dans les mesures de coordonnées pour des points bien définis. *L'exactitude sémantique* mesure la fidélité de données qui ne sont pas reliées à une position, telles que les classifications. *La cohérence logique* fait référence aux relations internes attendues dans la base de données. Elle peut être vérifiée sans faire appel au *monde réel*. *La complétude* porte sur l'exhaustivité d'une collection d'objets. L'évaluation de la qualité doit désormais faire appel à ces cinq composantes (Chrisman, 2005).

Enfin, nous assistons à l'émergence d'un nouveau concept appelé *terrain nominal*. Ce concept indique qu'un test d'exactitude ne se fait pas de manière aléatoire et que la mesure de l'erreur ou de l'exactitude ne peut être immédiate. Vous ne vous rendez pas sur le terrain sans avoir les filtres d'une approche établie permettant de classifier les objets, décrire leurs relations et mesurer leurs propriétés. Le test n'est pas effectué dans le monde *réel*, mais dans le monde *nominal* dans lequel les objets sont définis en fonction des spécifications et de techniques de mesure définies. Cela nécessite de faire une abstraction du monde réel et de définir au préalable le terrain nominal avec ses spécifications, ses tests et seuils adaptés lors d'évaluation de la qualité. Nous mesurons l'exactitude qui est la vérité à travers un terrain nominal. On assiste alors à l'apparition du terme *précision*. Ce dernier exprime à

quel degré la valeur produite s'écarte de la valeur attendue issue d'un terrain nominal ou d'un jeu de données dit de *référence* dont on connait la qualité.

Servigne et al. (2005) apportent plus de détails pour distinguer les deux concepts *précision* et *exactitude*. Les auteurs définissent la précision comme étant la résolution de mesure d'un phénomène avec instrument de mesure donné ou d'une méthode. De ce fait, la précision est fixe dès le début de la prise de mesure. Par contre, une exactitude évoque l'incapacité à mesurer la vérité d'un phénomène et admet l'existence des erreurs de mesures à chaque prise de mesure.

L'erreur correspond à un écart entre deux jeux de données parfois appelé une distance. Elle correspond à la notion d'exactitude. Par contre l'incertitude, elle apparait lorsqu'on émet un doute sur notre conceptualisation du monde réel face à un phénomène considéré. Elle fait référence à la notion de précision (Fisher et Tate, 2006).

Plus tard, Girres (2012b) reprend la définition faite par AFNOR(1955) en distinguant l'erreur accidentelle et l'erreur systématique qui correspondent respectivement à l'erreur aléatoire (estimée par le biais ou l'erreur moyenne) et l'écart-type (estimée par l'écart-type de l'erreur). Ces deux types d'erreurs vont exprimer l'inexactitude de mesures tandis que la notion d'incertitude se cadre dans un concept plus général *d'imperfection* au même titre que que les notions : *imprécision* et *l'incomplétude* (Batton-Hubert et al., 2019). Le caractère imprécis est exprimé par la difficulté d'exprimer une réalité alors que l'incomplétude traite l'absence de connaissances sur une réalité.

Plus précisément, Fisher et al. (2005) détaillent que tout processus utilisant les données géographiques prend en compte cette notion d'incertitude des données géographiques, et que cette incertitude est principalement liée au processus d'abstraction du monde réel. Ainsi, les auteurs distinguent les objets *bien définis* (par exemple un bâtiment dont ses frontières sont nettes et bien définis) des objets *mal définis* (par exemple une vallée ou encore une forêt dont la frontière dépend de la définition de la frontière). De cette différentiation, les auteurs distinguent trois formes d'incertitude à savoir *l'erreur*, *le vague* et *l'ambigu*, et proposent de les gérer à l'aide de théories mathématiques traitant du domaine de l'incertain. *l'ambigu* se manifeste lorsqu'on a un doute sur la manière de classifier un phénomène, il se subdivise en *désaccord* (sur les limites frontaliers par exemple) et en *non-spécificité* (sur le manque de précision des relations spatiales).

En somme ces travaux tentent de définir la qualité avec des terminologies différentes face à des contextes liés à la nature des données, aux méthodes et aux appareils de mesure utilisées. De nos jours ces terminologies constituent des imperfections dont les chercheurs convergent vers leur acceptation en tant qu'incertitude (Batton-Hubert et al., 2019). Nous estimons que ce terme englobe tout erreur et inexactitude qui s'entache aux données quelque soit leur source. Comme actuellement, l'évaluation de la qualité d'un jeu de données se fait à travers des spécifications et avec un jeu de données de référence, nous admettons que le terme *incertitude* pourrait renfermer tout écart susceptible d'exister entre un jeu de données pris comme référence dont on connait sa qualité et un autre jeu de données dont on souhaite le qualifier à travers des indicateurs de comparaison. Cet écart exprime la précision du jeu de données à qualifier par rapport au jeu de données de référence. Désormais l'étude de la qualité d'un jeu de données s'exprime par la précision vis-à-vis d'un jeu de données de référence et des spécifications quand elles existent.

3.1.2 La qualité selon la norme ISO 19154 :2014

En s'inspirant de la norme ISO 19154 :2014, on distingue les paramètres ou critères de qualité quantitatifs offrant une expression qualitative de la qualité et les critères dits qualitatifs. Les critères quantitatifs sont : *précision précision de position*, *précision thématique*, *précision temporelle*, *cohérence logique* et *complétude*. Un dernier critère qualitatif sur la qualité est dit *utilisabilité*.

Par ailleurs, certains auteurs préfèrent ajouter aux critères de qualité ,la précision sémantique qui se réfère à la nature des objets géographiques afin d'éviter la confusion avec la précision thématique faisant référence aux attributs des objets géographiques. Également dans la littérature on trouve d'autres termes : exhaustivité pour la complétude, précision géométrique avec la précision de position. La précision de position et la précision géométrique forment la précision spatiale.

La précision spatiale fait référence au degré de proximité entre une mesure d'une quantité et la valeur réelle acceptée de cette quantité. La précision positionnelle peut être absolue ou relative par rapport rapport au système de coordonnées de référence représenté alors que la précision géométrique dépend directement de moyens d'acquisition et de traitement de mesures (nombre de sommets à la saisie et algorithmes pour l'interpolation). La précision thématique fait référence à la fidélité aux noms de lieux , à l'exactitude de la classification et se calcule sur d'autres attributs (exemple : la heure des bâtiments, le nombre de voie, la hauteur des arbres, etc). La précision sémantique quant à elle, exprime la conformité de la nature des objets. La complétude s'évalue suivant les omissions et les commissions existantes selon le terrain nominal et le jeu de données évalué. La cohérence logique permet de vérifier si les objets décrits dans une base de données respectent d'abord la réalité (terrain nominal) puis la topologie et les relations spatiales entre eux. La précision temporelle nous renseigne sur la fraicheur de données en termes de date de prise de données, de période de validité du lot de données ainsi que la cohérence temporelle. L'usage décrit le cadre d'utilisation prévu pour un jeu des données.

Selon une analyse axée sur les paramètres, les travaux de Hangouët (2005) subdivisent la qualité en : la qualité interne et la qualité externe. En effet la qualité interne mesure le degré de conformité des paramètres intrinsèques à l'objet ou aux modèles ou à la conceptualisation issue du terrain nominal. Ce sont la précision positionnelle, géométrique, thématique, l'exhaustivité et la cohérence topologique/sémantique, etc. Quant à la qualité externe (fitness for use), elle se définit par la l'adéquation d'un jeu de données à une utilisation donnée. Elle fait appel à l'usage, la cohérence temporelle ainsi que des seuils d'acceptabilité sur les indicateurs de la qualité interne.

3.2 Évaluation de la qualité des données géographiques

L'évaluation de la qualité des données géographiques passe d'abord par définir une approche à travers on serait capable de mesurer la qualité. Cette dernière est évaluée à travers une une caractérisation puis une mesure de l'incertitude liée aux données et aussi à travers une étude de sa propagation.

3.2.1 Démarche d'évaluation d'un jeu de données

Plusieurs chercheurs ont tenté d'évaluer de manière précise un jeu de données à travers les paramètres de qualité ou parfois avec des tests statistiques ou encore à travers des algorithmes et des modélisations.

Au fait, la norme ISO 19114 (Jakobsson et Giversen, 2007) définissent des procédures pour évaluer la qualité des données géographiques. Elle propose deux méthodes d'évaluation de la qualité dont le résultat peut être quantitatif ou juste limité à une indication de la conformité des données vis-à-vis d'une spécification des produits :

- une méthode directe qui consiste à comparer un jeu de données avec d'autres données dites de référence. Cette comparaison se fait selon des spécifications formulées à travers un terrain nominal fait auparavant et selon bien sûr un ou plusieurs critères de qualité. (critère quantitatif) ;

- une méthode indirecte qui consiste à déduire ou à estimer une mesure de la qualité des données à partir des métadonnées.

Quelle que soit la méthode, l'évaluation porte sur l'ensemble du jeu de données ou par échantillonnage sur un sous-ensemble représentatif de la sélection. Comme déjà mentionné, l'évaluation de la qualité ou de l'incertitude ne se fait pas directement sur les données réelles à tester mais l'on agit à travers une abstraction du monde réel (terrain nominal) cherchant à mesurer la qualité suivant des spécifications, la vérité en soit n'étant pas inaccessible.

Dans le temps, les chercheurs ont tenté d'évaluer la qualité des données dites VGI (Volunteered Geographic Information) issues de contributions des utilisateurs du web. Parmi les premiers travaux qui se sont intéressés à évaluer des jeux de données nous pouvons citer ceux de Girres et Touya (2010). Les auteurs ont évalué un jeu de données OpenStreetMap en France en les comparant avec des données de référence produites par l'IGN afin afin de les qualifier. L'évaluation s'est basée sur les indicateurs de la précision spatiale et sémantique ainsi que l'exhaustivité. Aussi, les travaux de Siebritz (2014) ont conseillé d'évaluer la conformité des données OSM avec les données topographiques existantes de l'agence nationale de cartographie (Afrique du Sud) par le biais d'un seuil d'acceptabilité afin de ne retenir que les zones dont les données OSM respectent ce seuil. Par la suite, Siebritz (2014) suggère d'utiliser ces données-là soit pour détecter les nouveaux changements sur les villes et ainsi programmer des enquêtes de collecte de données pour la mise à jour, soit directement les intégrer dans la base en les soumettant à des corrections d'erreurs. L'évaluation de la qualité a porté sur la précision géographique.

3.2.2 Évaluation de l'incertitude

Dans un monde en évolution rapide, on a besoin de répondre finement à certains besoins et il s'avère que les moyens technologiques actuels le permettent surtout via le web 2.0. La qualité de l'information géographique n'est cependant pas garantie, les données qui en résultent sont susceptibles d'erreurs. Ainsi, les recherches et les développements doivent être orientés vers des mesures d'incertitude, une propagation d'erreurs(plus exactement

propagation d'incertitudes) et une évaluation de la qualité de l'information géographique.

Évaluer l'erreur impliquerait l'existence de la valeur réelle de la quantité. Or, cette vérité n'existe que dans des intervalles probables et constitue une hypothèse inaccessible (Li et al., 2012). Il est plus judicieux d'utiliser le terme *incertitude* qui sous-entend une certaine relativité sur la vérité et sera plus adapté pour transmettre le caractère vague (Burrough et Frank, 1996; Zhang et Goodchild, 2002).

Afin de mesurer l'incertitude, Fisher et al. (2005) proposent d'analyser d'abord la nature de l'incertitude. Ici, il est question de connaitre la nature de la *classe d'objets* à traiter (exemple : les sols), l'objet individuel à observer (exemple : unité de sol A), les propriétés à mesurer (exemple : description d'un champ) et enfin le traitement de ces observations (exemple : taxonomie numérique). Ainsi, les auteurs distinguent d'une part le cas où la classe d'objets et l'objet individuel sont bien définis alors l'incertitude est causée simplement par des erreurs et elle est de nature probabiliste, et d'autre part le cas où l'un de deux est mal défini. Dans ce cas-ci, on distingue encore deux cas :

- si l'incertitude est due à une définition *imparfaite*, alors cette définition est de type *floue* et peut être traitée comme un *sous-ensemble flou* ;
- si l'incertitude est due à une *ambigüité*, telle qu'une mauvaise classification, alors cela peut prendre à son tour deux formes (Klir et Yuan, 1995) :
  1. soit la classe (ou l'individu) est clairement définie, mais peut être à associée à deux, ou plusieurs classes différentes, on parle de *désaccord* ;
  2. soit l'assignation d'un objet à une classe est elle-même sujette à caution et c'est un problème de *non-spécificité* (Fisher et al., 2005) ;

La figure 3.1 résume un modèle conceptuel de l'incertitude. Bien évidemment, cette liste conceptuelle ne peut être tenue pour définitive. Et les définitions détaillées de ces différents concepts sont illustrées dans la littérature. Toutefois nous insistons sur le fait qu'il est nécessaire de définir l'incertitude, les objets à traiter et leurs caractéristiques face à chaque phénomène géographique. Ainsi, nous saurions caractériser l'incertitude dans sa globalité. Bien qu'on a mentionné au début du chapitre que l'incertitude représente le doute sur la conceptualisation d'une base de données à partir du monde réel. Ici, elle porte sur la mesure de l'information géographique et non sur la conceptualisation. On analyse la nature des données afin d'adopter une démarche de calcul de l'incertitude.

Par la suite, arrive la phase de modélisation de l'incertitude. Cette dernière cherche à rapprocher la distribution de la série des écarts à une distribution théorique afin de déterminer une fonction de densité (ou fonction de répartition) pour notre série des écarts. Ainsi, on pourrait être en mesure d'estimer un écart-type sur les écarts. Cet écart-type correspond à l'incertitude recherché pour un indicateur de qualité donné.

Dans cette thèse, si l'on souhaite caractériser des objets de type *bâti*, selon la classification faite par Fisher et al. (2005), ces données rentrent dans la catégorie des objets *bien définis* dont la mesure de leur incertitude doit faire appel à l'utilisation des méthodes probabilistes. Or dans notre contexte, la vérité terrain de nos données est inaccessible. Pour évaluer la qualité de notre jeu de données, nous procédons par une comparaison dudit jeu de données avec un jeu de données de référence sous l'admission que le jeu de données pris comme référence serait considérer comme une approximation des données de

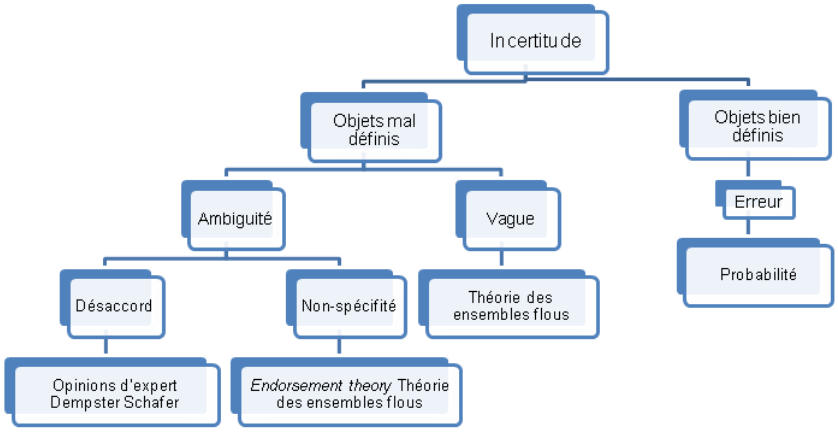


FIGURE 3.1 – modèle conceptuel de l'incertitude dans les données spatiales. Source : d'après Fisher (1999), adapté de Klir et Yuan (1995) tiré de Fisher et al. (2005)

vérité. Du fait aussi que nos données pourraient être entachées des divers erreurs dont les sources pourraient provenir des instruments de mesures, des méthodes de mesure ainsi que la nature stochastique du phénomène à mesurer, nous estimons que le terme adéquat pour qualifier notre jeu de données est celui de *incertitude*. Et l'écart du jeu de données à évaluer, mesuré par rapport au jeu de référence s'exprimerait l'incertitude de notre jeu de données. Cette incertitude combine des erreurs produites nos seulement lors de la saisie de la donnée mais aussi de leur propagation dans le processus de mesure des données.

3.2.3 Propagation d'erreurs

Une mesure est toujours entachée d'une erreur, dont on estime l'intensité par l'intermédiaire de l'incertitude. Lorsqu'une ou plusieurs mesures sont utilisées pour obtenir la valeur d'une ou de plusieurs autres grandeurs (par l'intermédiaire d'une formule explicite ou d'un algorithme), il faut savoir, non seulement calculer la valeur estimée de cette ou ces grandeurs, mais encore déterminer l'incertitude ou les incertitudes induites sur le ou les résultats du calcul. En fait, on veut déterminer de quelle manière chacune de ces incertitudes se répercute sur la grandeur finale. On parle de propagation des incertitudes ou souvent, mais improprement, de propagation des erreurs. Cette détermination est aussi appelée analyse de sensibilité d'une méthode.

Prenons un cas formel : on dispose d'une fonction f qui transforme un ensemble $\mathbf{x}$ de données en un produit $\mathbf{y} = f(\mathbf{x})$. Par exemple, $\mathbf{x}$ peut être une collection de modèles numériques de bâtiments et $\mathbf{y}$ peut représenter l'aire visible depuis un point d'observation donné, comme illustré sur la figure 3.2.

Cependant, en général, les données $\mathbf{x}$ utilisées en pratique sont entachées d'erreurs, et il est important de contrôler l'impact de cette erreur sur le résultat $\mathbf{y}$: on parle d'*analyse de sensibilité* de l'application f (Saltelli et al., 2000). Connaître l'influence du bruit de $\mathbf{x}$ sur une application f est primordial pour (1) quantifier l'incertitude sur le résultat obtenu à l'issue de l'application, (2) déterminer les paramètres et données critiques sur lesquels

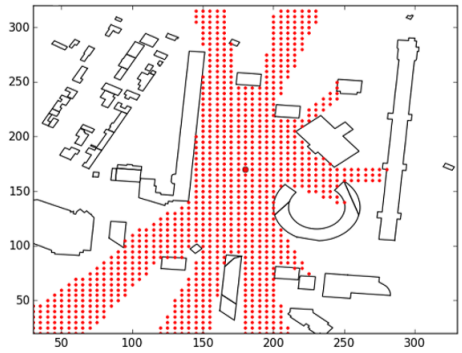


FIGURE 3.2 – Zone visible depuis un point d'observation. Données IGN BD TOPO© sur la commune d'Aubervilliers (93). La surface visible est évaluée à 22 275 m^2.

on doit investir le plus d'efforts et (3) déterminer la plage de fonctionnement de f, *i.e.* spécifier la qualité minimale des données utilisées en entrée garantissant le fonctionnement nominal de f.

Il existe dans la littérature trois grandes classes de méthodes pour propager l'erreur de $\mathbf{x}$ à f(x). Il s'agit de la méthode de la différentielle totale, de la méthode de propagation des variances et de la méthode de simulation dite de Monte Carlo.

- **méthode de la différentielle totale**
  Cette méthode consiste à déterminer les amplitudes de variation $\Delta\mathbf{x}$ sur les entrées, puis à les transférer sur Δf à l'aide d'une différentielle totale :

$$\Delta f = \Delta x_1 \left|\frac{\partial f}{\partial x_1}\right| + \Delta x_2 \left|\frac{\partial f}{\partial x_2}\right| + ... + \Delta x_n \left|\frac{\partial f}{\partial x_n}\right| \tag{3.1}$$

  où Δx_i désigne l'intervalle de variation de la i-eme composante des données $\mathbf{x}$.

  Cette méthode, simple à mettre en œuvre, nécessite toutefois que $\mathbf{f}$ soit suffisamment régulière pour être approchée par une linéarisation au voisinage du point de fonctionnement. Par ailleurs, une de ses limites majeures est de ne fournir qu'un intervalle absolu Δf, i.e. une plage de l'ensemble des valeurs potentiellement prises par $\mathbf{f}$ sans tenir compte de leurs probabilités d'apparition respectives, et va donc fournir en général des intervalles très larges et pessimistes.

- **méthode de propagation des variances**
  Une autre méthode consiste à utiliser le théorème de propagation des variances (pour le cas où f est linéaire) :

  Théorème 1 (Propagation des variances)**.** *Soit* $\mathbf{X} \in \mathbb{R}^n$ *un vecteur aléatoire de matrice de covariance* Σ_X *et* $\mathbf{A} \in \mathbb{R}^{m\times n}$*. Alors, la matrice de covariance* Σ_Y *sur* $\mathbf{Y} = \mathbf{AX}$ *est :*

$$\Sigma_Y = \mathbf{A}\Sigma_X\mathbf{A}^T$$

En particulier, en posant $A = [1, 1]$, on retrouve directement l'expression de la variance d'une somme de 2 variables aléatoires corrélées : $\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}(X, Y)$

Que faire lorsque l'application n'est pas linéaire ? Soit f une fonction scalaire non-linéaire, et X une variable aléatoire de variance $\text{Var}(X)$. Pour évaluer la variance de $Y = f(X)$ une solution prête à l'emploi consiste à linéariser f autour d'une valeur de référence a :

$$f(X) = f(a) + f'(a)(X-a) + \frac{f^{(2)}(a)}{2}(X-a)^2 + \ ... \ + \frac{f^{(n)}(a)}{n!}(X-a)^n + o\big((X-a)^n\big)$$

Alors, par les propriétés de la variance, et en se limitant à un développement limité à l'ordre 2, on a :

$$\text{Var}[f(\mathbf{X})] \approx \text{Var}[(f(a)+f'(a)(X-a)] = \text{Var}[f(a)]+f'(a)^2\text{Var}[X-a] = f'(a)^2\text{Var}[X])$$

On peut généraliser cette méthode pour une fonction $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, dont on peut calculer la matrice jacobienne $\mathbf{J}$ de terme général $(\mathbf{J})_{ij} = \frac{\partial f_i}{\partial x_j}$, où $f_1, f_2, ...f_m$ sont les composantes de f.

Théorème 2 (Propagation des variances : cas non-linéaire)**.** *Soit $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, une fonction régulière de matrice jacobienne $\mathbf{J}$ et $\mathbf{X}$ un vecteur aléatoire de $\mathbb{R}^n$ de matrice de covariance Σ_X :*

$$Var[f(\mathbf{X})] = \mathbf{J}\Sigma_X\mathbf{J}^T$$

$$\text{Var}[f(X)] \approx \text{Var}\big[f(a)+f'(a)(X-a)\big] = \text{Var}\big[f(a)\big]+f'(a)^2\text{Var}[X-a] = f'(a)^2\text{Var}[X]$$

Prenons le cas de $f : x \mapsto x^2$, et supposons que l'on travaille au voisinage de $a = 1$. On a alors :

$$\text{Var}[X^2] = (2a)^2\text{Var}[X] = 4\text{Var}[X]$$

Dans le cas où **f** est fortement non-linéaire (notamment avec des effets de seuil) alors la linéarisation de **f** n'est pas satisfaisante et la variance de sortie n'est plus réaliste. C'est le cas dans l'exemple donnée en figure 3.2, l'inter-visibilité est un phénomène binaire (une zone est visible ou ne l'est pas mais il n'y a pas de situation intermédiaire) donc typiquement non-linéaire. Par ailleurs, lorsque f n'a pas d'expression analytique (c'est le cas ici aussi, et en général dans toutes les applications où f est calculée avec un algorithme complexe), ces deux premières méthodes d'analyse de sensibilité sont inopérantes.

- **méthode de Monte Carlo**
  La méthode est fondée sur des simulations. Le principe consiste à introduire un bruit dans l'entrée **x**, et d'observer l'impact sur f(x) pour un grand nombre de réalisations. Ainsi il est possible d'évaluer tous les indicateurs statistiques souhaités sur la population de sortie (biais, écart-type, erreur RMSE, percentiles, valeurs extrêmes...). On peut réitérer l'expérience pour différents niveaux de bruits en entrée. Par exemple, sur la figure 3.3, nous ajoutons un bruit sur les coordonnées des bâtiments (d'écart-type $\sigma = 5$ m sur la rangée d'images supérieure, et $\sigma = 10$ m sur la rangée inférieure) et nous étudions l'impact sur la surface visible calculée.

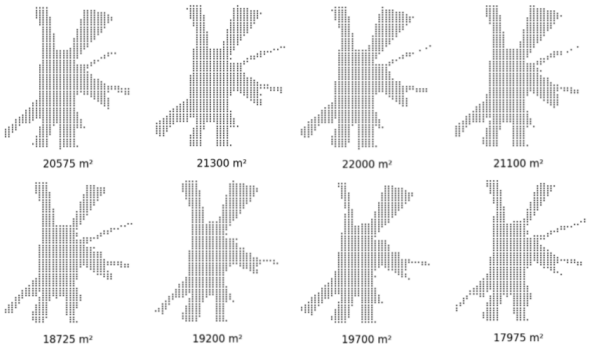


FIGURE 3.3 – Analyse de sensibilité de la surface visible en fonction de la précision des bâtiments.

Les résultats montrent un RMSE de 1304 m^2 (soit environ 6% de l'aire totale) sur la surface visible pour une précision de 5 m en entrée. Cette valeur passe à 2484 m^2 (11% de l'aire totale) pour $\sigma = 10$ m. Dans les deux cas, le biais d'estimation est négatif et représente environ les deux tiers de l'erreur total, ce qui laisse penser que le bruit sur les données induit plutôt une erreur systématique tendant à sous-estimer l'aire visible.

On pourra trouver une étude similaire complète dans les travaux de Biljecki et al. (2015) qui étudient l'impact de l'erreur du bâti sur l'évaluation du potentiel solaire.

Malgré l'apparente simplicité de cette méthode, il faut prêter attention à ce que le bruit introduit soit représentatif de l'erreur typique entachant les données d'entrée. L'utilisation d'un bruit blanc (*i.e.* non-corrélé) pour perturber l'entrée **x**, peut mener à une sous-évaluation, ou au contraire à une sur-évaluation de l'impact sur $f(\mathbf{x})$. En particulier, quand f est plutôt de type additive (sommes, cumuls, moyennes, intégrations...) le bruit blanc a une tendance systématique à sous-estimer l'impact. À l'inverse, une fonction de type soustractive (différentielle, pente, vitesse, accélération, écart...) aura tendance a être plus sensible au bruit blanc qu'à l'erreur typique entachant les données en réalité. Dans l'exemple 3.3, le bruitage indépendant des coordonnées des sommets n'est pas une solution satisfaisante, et pour deux raisons principales : en premier lieu, le bruit blanc sur les coordonnées a tendance à facilement *bloquer* les lignes de vues. Dans la pratique, les erreurs entâchant les bâtiments sont spatialement corrélées, et les lignes de vues ont globalement tendance à être conservées (au moins jusqu'à la portée de corrélation des erreurs). Deuxièmement d'un point de

vue plus pragmatique, l'introduction d'un bruit indépendant sur des données vectorielles (polylignes, polygones...), en plus de produire des formes irréalistes, peut résulter en des erreurs topologiques (auto-intersections, chevauchements...), risquant ainsi de rendre plus difficile, voire impossible le calcul de f (si celle-ci n'est plus bien définie).

3.3 Qualification des données géographiques volontaires

3.3.1 Les données géographiques volontaires, leur intérêt et leur qualification

Jusqu'à ces dernières décennies, cartographier un contenu géographique était un travail d'organismes officiels et des personnes qualifiées d'expertes et disposant des outils fiables pour relever des objets ou phénomènes géographiques. Les utilisateurs exploitaient ces données à travers des plateformes cartographiques. Cependant, avec l'avènement du Web 2.0, ces utilisateurs peuvent contribuer ou générer des contenus géographiques, et produisent ce qu'appelle pour la 1ère fois Goodchild (2007) de l'information géographique volontaire (Volonteered Geographic Information - VGI). Il s'agit d'un grand nombre de citoyens, ayant peu de qualifications caractérisés par leur volontariat, qui produisent un grand volume de données, donnant naissance à des plateformes de contenus collaboratifs tels que Wikipedia ou OpenStreetMap. Une des caractéristiques de ces données est que l'on ignore la qualité a priori.

Dès lors, nous avons assisté à une croissance spectaculaire des données VGI, notamment OpenStreetMap qui attire l'attention des chercheurs, qui cherchent à évaluer la qualité de ces données pour pouvoir bénéficier de leur richesse en vue de mettre à jour des référentiels cartographiques ou parfois intégrer les données dans des processus de fusion de données environnementales dans des systèmes d'information géographique (Yan et al., 2020; Bordogna et al., 2016; Garcia-Martí et al., 2017).

La définition de la qualité de données géographiques évolue. Autrefois, la qualité était évaluée à travers les indicateurs des normes ISO, mesurant ainsi l'écart d'un jeu de données avec un jeu de données de références dont les spécifications sont précisées, dans un cadre où les données sont collectées et produites avec un effort commandé et une méthode spécifiée et uniforme. Ce qui conférait à cette qualité le vertu d'être supposée homogène et cohérente (Van Exel et al., 2010).

De nos jours, les collections VGI sont caractérisées par une qualité hétérogène et diverse, du fait qu'elles sont collectées en utilisant différentes méthodes (par exemple, les traces GPS, la numérisation d'images) et par différentes personnes avec des motivations et des préférences différentes (Yang et al., 2018; Minaei, 2020). De plus, les contributeurs et les contributions ne sont pas répartis de manière égale dans l'espace (Ma et al., 2015; Viana et al., 2019).

Ainsi, Van Exel et al. (2010), introduisent le concept de *qualité de foule* (Crowd Quality) pour caractériser la qualité des données spatiales participatives. Crowd Quality (CQ) tente de quantifier *l'intelligence collective de la foule générant des données* dans un contexte spatio-temporel. L'auteur distingue les aspects de la qualité liés à l'utilisateur (contributeur) et ceux liés au contenu collaboratif. Les aspects de la qualité liés au contenu

collaboratif peuvent correspondre aux éléments (indicateurs) de qualité définis par ISO, étendus avec des éléments de qualité spécifiques aux données de crowdsourcing.

La définition de la qualité liée au contenu collaboratif est complétée par Bordogna et al. (2014) en distinguant :

- La composante sémantique de l'entité géographique : ce sont des descriptions textuelles et des annotations, des mots-clés, catégories et autres attributs spécifiques dont les valeurs peuvent être obtenues par des mesures de certaines propriétés de l'entité. Ceux-là décrivent ce qu'est l'entité. Par exemple, pour représenter l'incertitude des connaissances sur la localisation d'une entité, il a été calculé une valeur correspondant à une distance maximale depuis le point de localisation des entités géographiques. Cette valeur constitue un attribut descriptif de l'entité.
- La composante contextuelle : elle décrit la généalogie de l'information, comprenant le nom de l'auteur et de l'organisation, indiquant qui a créé ou observé l'entité ; les empreintes géographiques spécifiant où l'entité a été observée, les dates indiquant quand l'entité a été observée et/ou créée.

De façon détaillée et exhaustive, Criscuolo et al. (2016) classent les indicateurs en fonction de trois types de qualité en s'inspirant de l'ISO 19113-15 et de la littérature :

- Qualité intrinsèque : elle correspond à la qualité interne d'ISO, et ne dépend que des caractéristiques de l'information géographique volontaire (précision spatiale, précision temporelle, précision sémantique ...etc.). Il est ajouté dans cette catégorie, l'indicateur de l'intelligibilité qui mesure que la contribution peut être comprise.

- Qualité extrinsèque : elle dépend des caractéristiques du contexte et répond aux besoins d'évaluer à la fois la crédibilité de l'auteur et de la fiabilité ou la véracité de l'information elle-même. ((Flanagin et Metzger, 2008) ; (Galloway et al., 2006; Genet et Sargent, 2003).

- Qualité pragmatique : elle correspond à la qualité externe de l'ISO et mesure la capacité à répondre aux besoins d'un utilisateur donné et/ou d'un usage.

Par ailleurs, l'évaluation de la qualité interne (selon ISO) des données VGI nécessite de disposer un jeu de données de référence. Ainsi, Antoniou et Skopeliti (2015) distinguent des **mesures de qualité** et des **indicateurs de qualité** respectivement dans le cas où on compare avec des données de référence et dans le cas où on ne dispose pas de telles données.

En s'appuyant sur ces travaux bibliographiques, nous définissons le vocabulaire de notre thèse de la manière suivante :

- Qualité extrinsèque : cette qualité est évaluée avec ou à travers une source externe à la donnée géographique volontaire.

    1. Comparaison avec base de référence : on évalue la qualité des données VGI par comparaison avec des données prises comme référence et dont on suppose

qu'elles sont de qualité connue. L'évaluation se fait à travers les mesures de qualité extraites des normes ISO 19517(précision géométrique, précision de position, précision thématique/attributaire, exhaustivité, comparaison d'itinéraires, etc.). Les données de référence sont en général des données d'autorité appartenant à des agences nationales de cartographie.

2. Évaluation de la crédibilité des contributeurs : on évalue la qualité des données VGI à travers les contributeurs à l'aide des indicateurs de qualité qui évaluent la confiance, la réputation, expérience/expertise, connaissance locale, la fiabilité, etc (Flanagin et Metzger, 2008).

- Qualité intrinsèque : elle s'appuie sur la donnée en elle-même, soit en prenant en compte l'état présent de la donnée elle-même, soit l'historique de la donnée, ou encore par en prenant en compte la cohérence avec les données de son voisinage (contexte spatial).

  1. Qualité intrinsèque basée sur les données elles-mêmes : on définit des indicateurs intrinsèques en se basant uniquement sur les caractéristiques de la donnée (forme, position, granularité). Parmi les indicateurs intrinsèques, on cite : la résolution géométrique, la granularité et la cohérence topologique. Dans un travail qui porte sur la détection des incohérences, Touya et Brando-Escobar (2013) ont proposé des indicateurs intrinsèques permettant de mesurer la résolution géométrique et la granularité des entités géographiques provenant d'OSM. La résolution géométrique se mesure à travers la densité des sommets, la longueur d'arête médiane (longueur des arêtes entre les sommets) tandis que la granularité s'exprime par la taille des entités et le contour le plus court (longueur de bord entre 2 sommets).

     Bien que la cohérence logique soit évaluée majoritairement à travers des spécifications (par exemple le schéma conceptuel), elle peut être considérée comme un **indicateur de qualité** à travers sa composante topologique dite **cohérence topologique**. La cohérence topologique examine l'exactitude des caractéristiques topologiques et détecte les incohérences topologiques.

     En guise d'exemple, Hashemi et Abbaspour (2015) détectent de erreurs de dépassements, des doublons, des zones non-fermées et des polylignes brisées.

     En guise d'exemple, Neis et al. (2012) observent des erreurs topologiques telles que :

     (a) des routes ne sont pas connectées par erreur aux jonctions ;
     (b) des routes qui se croisent sans avoir des nœuds en commun ;
     (c) des duplications des géométries des routes.

  2. Qualité fondée sur l'historique de la donnée : il s'agit ici de s'appuyer sur l'historique d'édition pour évaluer la qualité des données. On utilise comme indicateurs intrinsèques sur l'historique, le nombre d'éditions de l'entité géographique, le nombre moyen de tags par entité, le nombre total des contributeurs actifs sur

l'entité géographique, l'évolution dans le temps de la longueur du réseau OSM sur une zone, etc (Barron et al., 2014).

3. Qualité sur le contexte spatial : il s'agit d'évaluer la qualité des données à travers la cohérence du thème d'un objet géographique avec la localisation ou la comparaison entre deux thèmes d'un même objet. Nous pouvons comparer par exemple, les bâtiments avec l'occupation du sol définie dans OSM, ou les entrées des bâtiments avec les emplacements de ces bâtiments (écarts) ou encore voir si les arrêts de bus s'alignent le long des routes.

Toutefois pour assurer ou évaluer la qualité des données VGI (en particulier celles d'OpenStreetMap), Goodchild et Li (2012) ont proposé trois approches :

- Le crowdsourcing ou l'approvisionnement par la foule : ce concept se fonde sur l'idée que l'on peut trouver la solution d'un problème en le soumettant à un groupe sans tenir compte de la qualification des individus qui le composent, parfois mieux qu'en considérant l'avis d'un seul expert. Il confère à un groupe la capacité de valider ou corriger une contribution faite par un individu, et plus ce groupe s'élargit, meilleure est l'assurance sur la qualité, car les membres du groupe se corrigent et s'améliorent jusqu'à converger vers la vérité. Nous pouvons citer par exemple, les travaux de Haklay et al. (2010) qui ont montré que la précision de positionnement d'un objet géographique s'améliore à mesure que le nombre d'éditeurs augmente, mais qu'aucune amélioration supplémentaire n'est évidente lorsque le nombre d'éditeurs dépasse 13.

  Selon cette approche, pour évaluer la qualité d'OSM, nous devons se référer à l'ensemble des indicateurs intrinsèques basés sur l'historique de la contribution.

- L'approche sociale : des personnes de confiance ayant une certaine réputation par leur activité de contributeurs servent de garants pour maintenir et contrôler la qualité des autres contributions. En effet, en se fondant sur l'exemple de Mooney et Corcoran (2012b) dans le cadre d'une analyse des données OSM pour les îles britanniques, il a été constaté que pour les éléments fortement édités (au moins 15 fois par élément) 84% des éditions ont été faites par 12% des contributeurs. Cela montre bien qu'y a un petit nombre de contributeurs très actifs qui agissent comme modérateurs ou gardiens. Ce qui fonde l'approche sociale sur une hiérarchisation des personnes de confiance à travers une étude de crédibilité du contributeur (Flanagin et Metzger, 2008). Par exemple, une personne qui fait des contributions prolifiques et qui attirent peu d'éditions (au sens de correction) se voit attribuer une note élevée et se rapproche du statut de modérateur ou gardien. Truong et al. (2019a) modélisent les interactions sociales entre contributeurs à travers un graphe multi-couches contenant différents types d'interactions afin de définir des profils typiques.

  Selon cette approche, pour évaluer la qualité des données OSM, nous devons nous référer à l'ensemble des indicateurs extrinsèques portant sur les contributeurs de manière à leur attribuer une crédibilité, cette note représentant par exemple le degré de confiance, de réputation et de connaissance locale des contributeurs. Ce raisonnement confère différents statuts aux contributeurs, ce qui permet d'assurer la qualité

de leurs contributions ou de celles qu'ils vont valider.

- L'approche sur les lois de connaissance de la géographie : on utilise la 1ère loi de Tobler stipulant que les objets qui sont proches se ressemblent d'avantage que les choses lointaines. Cela implique qu'un événement sur un lieu ou la localisation d'un objet géographique sur un lieu doit être cohérent avec ce que l'on sait déjà sur le voisinage de ce lieu. Dans Goodchild et Li (2012), les auteurs citent l'exemple d'une cafeteria géoreferencée ponctuellement sur Flickr sur un plan d'eau, ce qui paraît assez improbable, cette localisation est donc en contradiction avec le contexte spatial (**parc**) en postulant qu'un commerce ne devrait pas se trouver dans une zone historique. Cette contribution semble être mal localisée et donc vouée à être rejetée.

  Selon cette approche, pour évaluer la qualité d'OSM, nous utiliserons des indicateurs se rapportant au contexte spatial de sorte à vérifier si le thème des contributions est en accord avec le lieu ou avec le thème du lieu.

3.3.2 Qualité par comparaison

Le choix de comparer un jeu de données issu de la base OSM avec un autre jeu provenant d'une base dite de référence n'est pas anodin. En effet, une base de données de référence est une base certifiée par une institution d'autorité dont sa mission principale est la production, le maintien, la diffusion et la protection des données géographiques servant de base commune à l'ensemble des acteurs agissant sur un territoire. Cette action confère à la base de données de référence, une certaine crédibilité et fiabilité. Cela justifie de comparer la base OSM à avec une de base de référence pour tester la qualité des données volontaires de la base OSM. Certains pays ont déjà évalué des jeux de données de la base OSM de leur pays à travers une comparaison des données d'autorité de leurs agences respectives. L'évaluation se fait toujours en utilisant des jeux de données sur des zones diversifiées soit en terme de paysage soit en terme de thématique (le réseau routier, le bâti). Ce sont entre autre l'Allemagne, le Royaume-Uni et la France.

L'évaluation passe par une étape d'appariement de données qui consiste à retrouver les objets homologues dans les deux bases de données (le jeu d'OSM et le jeu de données de référence). L'appariement peut être automatique ou manuel. Puis pour un couple apparié, on mesure l'écart de la donnée OSM à celle de la référence à travers les mesures de qualité extrinsèque définis auparavant. Nous détaillons par la suite l'ensemble des travaux qui ont procédé à l'évaluation de la qualité d'OSM par comparaison en fonction des mesures la qualité extrinsèque.

La précision de position : écart de la position des objets OSM par rapport aux objets homologues de la référence

Parmi les travaux qui ont évalué les données OSM à l'aide de jeux de données de référence officiellement recueillis, on cite les travaux de (Kounadi, 2009) et (Ather, 2009) qui ont comparé les routes de Londres d'OSM avec celles la base de *l' Ordnance Survey Meridian 2* en étudiant la proportion de chevauchement des routes dans une zone tampon servant de résolution. La précision positionnelle de l'OSM a été réalisée par une comparaison statistique et visuelle. La comparaison statistique a porté sur un échantillon de segments d'autoroutes tandis que la comparaison visuelle de la précision de la posi-

tion a été effectuée en utilisant 100 échantillons dans cinq tuiles d'OSM. Les résultats ont montré qu'en moyenne 80 % des lignes OSM se chevauchent avec les lignes de la base de référence. La distance totale moyenne (écart) des 100 échantillons s'est avérée être d'environ 6 mètres. Haklay (2010) renforcent cette étude et précisent que le pourcentage de chevauchement a atteint 85,80 % dans le Sud de Londres, 85,19 % dans le Nord, 81,03 % dans l'Ouest et 80,80 % dans l'Est de Londres.

Par ailleurs, Helbich et al. (2012) ont réalisé l'analyse statistique de la précision positionnelle de trois sources de données différentes pour l'année 2011, à savoir OSM, TomTom (TT) et les données d'enquête (SD), pour une ville allemande de taille moyenne bien cartographiée. L'algorithme de prétraitement a permis d'extraire 121 carrefours routiers identiques dans les trois ensembles de données. En tenant compte du fait que le SD sert d'ensemble de données de référence spatialement précis, l'écart spatial entre le SD respectivement avec les données d'OSM et celles du TT, a été évalué. L'écart spatial moyen s'est avérée être environ 1 mètre plus petit dans l'ensemble de données OSM par rapport à celles du TT tout en observant que la variance sur l'écart est significativement plus grande en données OSM qu'en données du TT.

En outre, Jackson et al. (2013) comparent un jeu de données du Projet Collaboratif OpenStreetMap (OSMCP) de l'US Geological Survey (USGS), constitué des POI de type ponctuel (sur des écoles) avec un jeu de données de référence. L'OSMCP représente une variante hybride avec la pratique de VGI dont les volontaires saisissent les données qui par la suite sont soumises au contrôle qualité par l'agence gouvernementale (l'USGS), l'objectif étant que l'USGS utilise ces données comme complément à leurs ensembles de données officielles. Les données de référence sur les écoles sont produites à partir d'une liste provenant des listes d'écoles publiques et privées du ministère de l'éducation. Après comparaison, l'écart spatial médian des écoles appariés entre OSMCP et la base de référence vaut 33 m. Bien que ce chiffre semble assez élevé, on précise qu'il s'agit des données ponctuelles qui ne sont pas toujours placées de façon consistantes.

Fan et al. (2014) se consacrent à une évaluation de la qualité des données sur les empreintes des bâtiments dans OSM. Ils comparent les empreintes des bâtiments OSM et les empreintes des bâtiments de la base de référence ATKIS (German Authority Topographic-Cartographic Information System) pour la ville allemande de Munich. La précision de la position en question évalue la relation entre la valeur des coordonnées d'un bâtiment dans OSM et la réalité sur le terrain. Dans le travail présenté, les points correspondants d'une paire d'empreintes de bâtiment dans deux ensembles de données sont trouvés en premier. Ensuite, la précision de la position est calculée comme étant la distance moyenne de ces points correspondants. Le décalage moyen entre les empreintes des bâtiments OSM et les empreintes des bâtiments ATKIS est de 4,13 m avec un écart-type de 1,71 m. Le décalage le plus important est de près de 15 m, tandis que le plus petit est inférieur à un centimètre. La distribution des décalages est proche d'une distribution normale.

D'autres chercheurs ont utilisé cette méthode de comparaison pour évaluer la précision géométrique de données OSM, en occurrence, les travaux de Girres et Touya (2010) qui ont procédé à apparier manuellement les objets OSM d'un réseau routier. Cette approche manuelle a été préférée par rapport à une approche automatisée pour éviter toute erreur de traitement (Senaratne et al., 2017).

Dans l'ensemble, les auteurs sont unanimes à l'idée qu'il existe une forte dispersion statistique des écarts spatiaux moyens. Cela révèle l'existence d'une forte hétérogénéité de la précision de position pour les données OSM (Girres et Touya, 2010).

Enfin, les travaux de Graser et al. (2014) ont mis en place deux indicateurs de performance mettant en évidence la qualité des réseaux des rues dans la région de Vienne. Il s'agit d'une comparaison faite entre les données OSM et le graphe de référence officiel autrichien à travers l'indicateur sur les restrictions de virage et l'indicateur sur les rues à sens unique.

Précision thématique et sémantique

Les attributs des données OSM sont souvent incomplets ou manquants et quand ils existent, ils ne sont pas toujours corrects (ce sont les données qui sont mal classées ou les valeurs des attributs quantitatifs erronées). Ce qui rend la précision thématique/sémantique assez faible en général. Mooney et Corcoran (2012a) estiment que la façon dont les contributeurs marquent ou annotent les objets dans OSM posent de sérieux problèmes en observant que les valeurs attribuées aux attributs **name** et **highway** sont souvent sujettes à des changements fréquents et inattendus. Ils justifient l'origine de ces erreurs dans OSM par une annotation manuelle faite par des contributeurs au lieu d'utiliser les valeurs des **tags** disponibles sur la plateforme permettant de saisir ces valeurs de manière **normalisée**. Ces problèmes pourraient avoir un effet potentiellement préjudiciable sur la qualité des données OSM tout en nuisant à la perception de l'OSM dans la communauté SIG.

Pour éviter cela, Senaratne et al. (2017) citent les travaux de Codescu et al. (2011), Vandecasteele et Devillers (2013), Ali et al. (2014) qui ont développé des méthodes d'appariement de similarité sémantique, évaluant automatiquement l'annotation des tags dans OSM en fonction de la signification sémantique de telles tags. Cela pourrait éviter par exemple des erreurs sémantiques provenant des mauvaises spécifications des routes comme témoignent les travaux de Girres et Touya (2010) après avoir observés que les routes classées comme **secondaires** dans l'ensemble de données de référence ont été classées comme **résidentielles** ou **intermédiaires** par les contributeurs dans les données OSM. Concernant ces inexactitudes, ces mêmes auteurs mentionnent l'absence d'une classification normalisée et de règles de dénomination. Bien qu'il existe des valeurs de **tags** disponibles sur la plateforme, il manque aussi des règles pour les utiliser. Tout cela autorise aux contributeurs à entrer des étiquettes et des valeurs qui ne sont pas présentes dans la spécification d'OSM.

Par la suite Vandecasteele et Devillers (2015) ont introduit un système de recommandation d'étiquettes pour les données OSM qui vise à améliorer la qualité sémantique des balises. Le plugin OSMantic pour l'éditeur Java OpenStreetMap peut suggérer automatiquement les tags pertinents pour les contributeurs pendant le processus d'édition.

A terme, les travaux de Yan et al. (2020) estiment que les recherches futures doivent renforcer le contrôle de la qualité avant la contribution plutôt qu'à leur rectification. Cela doit aboutir par trouver un compromis entre la liberté des contributeurs et la gestion de leur comportement en améliorant par exemple la conception des plateformes de saisie du VGI comme Poplin et al. (2017) l'ont déjà proposé.

Complétude :

Parmi les travaux qui ont porté sur la complétude, on relève ceux de Koukoletsos et al. (2012) proposant d'utiliser une méthode d'appariement automatique des entités linéaires à partir d'ensembles de données de référence et d'OSM. La méthode combine des contraintes géométriques et des contraintes d'attributs (nom et type de route) afin de traiter des ensembles de données hétérogènes, en tenant compte du fait que des attributs peuvent être manquants. Lors des tests sur les ensembles de données d'OSM et de ITN (Référence) pour les zones rurales et urbaines sélectionnées, les auteurs jugent que la mise en correspondance des données s'est avérée efficace, avec des erreurs de correspondance entre 2,08 % (zones urbaines) et 3,38 % (zones rurales). L'exhaustivité du VGI est ensuite calculée pour les zones plus petites (tuiles). Leurs résultats montrent que généralement la base OSM est plus compète que la base ITN en zone urbaine et moins complète en zone rurale. Ces résultats concordent avec d'autres obtenus au Royaume Uni et ailleurs sur la complétude. Toutefois, les auteurs précisent que la densité élevée en zones urbaines des données OSM renvoie généralement à des informations non pertinentes pour les spécifications de collecte des données ITN. Cela signifie que certains objets tronçons présents dans OSM en zone urbaine n'ont pas leur place dans la base ITN du fait de ses spécifications.

Par ailleurs Neis et al. (2012) étudient la complétude du réseau routier de l'Allemagne en Juin 2011 à travers une comparaison relative entre OSM et une base de données commerciales. Les résultats montrent que pour la navigation automobile, le réseau routier d'OSM est plus petit que celui de Tom-Tom (la base commerciale) d'environ 9% plus petit que celui de Tom-Tom (la base commerciale) bien que le réseau global d'OSM est environ 27 % plus grand que celui de Tom-Tom. Quant au réseau piétonnier, celui d'OSM s'avère plus grand à raison d'environ 31% que celui de Tom-Tom.

D'après l'étude faite en France par Girres et Touya (2010), il semble qu'en termes de nombre d'objets, le réseau routier français d'OSM est loin d'être complet pour tous types de thème confondus par rapport au BD TOPO (base de référence). Cependant, la longueur/surface totale du réseau routier OSM semble proche de celui de la base de référence. Les auteurs expliquent cela par le fait que les petits objets sont plus susceptibles de manquer dans OSM parce que les contributeurs se concentrent davantage sur la capture d'objets attrayants (qui sont les plus utiles pour leur intérêt). Il a été remarqué aussi, pour OSM, que les zones rurales ayant peu de contributeurs (qui sont probablement des zones à faible population) étaient moins couvertes comme l'ont témoigné également les travaux de (Haklay, 2010).

3.3.3 Qualité fondée sur les contributeurs

L'évaluation de la qualité des VGI semble être difficile quand on ne dispose pas de données de référence. Les chercheurs se basent sur l'approche sociale pour hiérarchiser les contributeurs en étudiant leur comportement en particulier les caractéristiques de leurs contributions et la manière dont ils interagissent avec les autres utilisateurs. Beaucoup des travaux tentent de les classer afin de dresser des typologies qui pourraient conférer aux contributeurs certaines caractéristiques. Puis, ils établissement des indicateurs sur les contributeurs en mesurant la crédibilité ou la fiabilité d'un contributeur pour en déduire une fiabilité ou une qualité sur sa contribution. Cela correspond à une mesure de la qualité extrinsèque de l'information géographique volontaire.

Une première étape d'analyse de comportement du contributeur semble être requise. En

se basant sur les travaux antérieures, notamment ceux de Neis et Zielstra (2014) ou encore Truong et al. (2019a) livrent une analyse synthétique du comportement des contributeurs selon trois catégories distinctes : leur participation, leur habitude à cartographier et leur motivation. La participation du contributeur s'analyse soit autour du nombre d'objets crées (Neis et Zipf, 2012), soit sur la durée d'activité du contributeur (dimension temporelle) (Begin et al., 2016), tandis que l'habitude ou style cartographique s'évalue au travers soit de sa préférence pour un type d'opération cartographique spécifique (sur la géométrie, sur les clés) (Mooney et Corcoran, 2012c) ou un type d'entité spécifique (routes, bâti) (Bégin et al., 2013), soit leur interaction avec un autre contributeur (la co-édition, en modifiant la géométrie de l'édition d'un autre contributeur), (Stein et al., 2015). Enfin, la motivation du contributeur s'établit soit autour de l'expertise du contributeur (Coleman et al., 2009; Kuai et al., 2016) soit sur son genre (Duféal et al., 2016; Truong et al., 2019a; Neis et Zipf, 2012).

À l'issue de l'analyse du comportement des contributeurs, l'enjeu majeur réside dans le fait de définir une démarche évaluant la qualité des données à travers la fiabilité des contributeurs. Parmi les travaux antérieures, Bishr et Janowicz (2010) proposent deux mesures de qualité à savoir la confiance et la réputation. La confiance du contributeur se mesure à travers la confiance accordée à ses contributions tandis que la réputation correspond à la perception de la faillibilité d'une personne par la communauté. C'est un degré de confiance à double sens. Si le contributeur se voit contredire par un groupe de personnes, il perd en réputation. Cela diminue sa cote de confiance pour la suite. Selon le modèle proposé par les auteurs, la réputation est propre à un contributeur, mais la confiance est propre à une contribution car normalement elle varie entre les contributions d'un même contributeur. Celle-ci dépend de la proximité du contributeur avec la contribution (de sa confiance), et de l'historique d'édition après cette contribution.

Par ailleurs, par analogie au modèle précédent, Lodigiani et Melchiori (2016), proposent un modèle de graphe récursif basé sur l'algorithme *PageRank* de Google personnalisé. Sur ce modèle les contributeurs et leurs contributions ainsi que leurs relations mutuelles sont modélisés comme les nœuds d'un graphe. Les nœuds des contributeurs ont des liens entrant (sa relation avec un autre contributeur) et des liens sortants (lien vers sa contribution) tandis que les nœuds de contributions ont seulement des liens rentrants issus de leurs contributeurs.

Dans l'algorithme du PageRank de Google, la pertinence (PageRank) d'une contribution B est déterminée de manière récursive par le nombre total de liens pointant vers elle où la valeur de chacun de ces liens (ou score) est proportionnelle au nombre total des liens pointant A, c'est-à-dire la pertinence de A (A étant un contributeur). Le PageRank est ici utilisé pour exprimer la réputation d'une contribution et /ou d'un contributeur. Il s'agit d'un processus de Markov où le PageRank correspond à la probabilité stationnaire d'une chaîne de Markov [1].

3.3.4 Qualité intrinsèque basée sur les données elles-mêmes, actuelles

Les indicateurs intrinsèques définis dans l'ISO ne peuvent pas être appliqués aux données d'OSM car leur processus de cartographie est différent de celui des données faisant autorité pour ces raisons : absence de spécifications, utilisation d'outils différents pour la

1. hhttps ://fr.wikipedia.org/wiki/PageRank

saisie des données et variété des profils des contributeurs. Certains indicateurs intrinsèques sur la donnée géographique, ont été formulés pour évaluer ces données (Antoniou et Skopeliti, 2015).

Des chercheurs se sont donc concentrés sur les indicateurs qui pourraient révéler la qualité des données en examinant uniquement les données elles-mêmes. Par exemple, Ciepłuch et al. (2011) ont utilisé la densité de nœuds dans une entité, comme indicateur de qualité pour analyser la qualité des données OSM. Bien qu'il s'agisse d'une simplification de problèmes d'analyse spatiale plus complexes, cet indicateur nous donne un aperçu simple de l'activité de la communauté OSM pour une région particulière. Sur une étude faite sur l'OSM de Lituanie, les résultats montrent que la cartographie de cette activité est concentrée autour des zones à forte population. Cette cartographie permet plus en détail d'afficher la densité de certains types d'éléments tel que des points d'intérêts (POI) ainsi que la fréquence d'apparition de balises spécifiques.

Enfin dans un contexte de détection des mises à jour dans les données d'autorité, Ivanovic et al. (2019) ont étudié des traces GNSS fournies par le principe du crowdsourcing par des contributeurs. Parmi les défauts de ces traces, on retient celui de l'hétérogénéité de la précision de position issue de certains points n'appartenant réellement pas au vrai chemin la trace GNSS. Ces points sont soit considérés des points aberrants provenant d'une fausse acquisition de coordonnées, soit des points secondaires constituant des localisations des activités secondaires des humains (nommé **comportement humain secondaire**-SHC) en dehors de leur chemin. Pour filtrer ces points, les auteurs ont entre autre formulé des indicateurs intrinsèques tels que l'élongation pour le SHC, la valeur moyenne des angles de changement des trois direction, le changement brusque de distance et vitesse entre deux points consécutives, la résolution spatiale d'une trace, la différence maximale d'élévation entre un point avec son précédent et le suivant, et enfin la moyenne de distances et vitesses pour un triplet des points.

3.3.5 Qualité basée sur l'historique

L'évaluation de la qualité à travers l'historique des contributions repose sur la force de crowdsourcing. Bien que les contributeurs diffèrent de par leur connaissance sur la zone et leur motivation, ils s'engagent dans un cadre dans lequel ils interagissent en se corrigeant et se confirmant à travers leurs contributions faites sur une zone donnée. En admettant le principe disant que les données qui ont subi des modifications successives par différents auteurs sont susceptibles d'être de meilleure qualité, l'étude de l'historique sur la contribution se voit désormais une bonne piste pour évaluer la qualité d'une donnée volontaire. Par la suite, on détaille les différentes manières d'évaluation de la qualité intrinsèque basé sur l'historique d'édition en fonction des indicateurs de la qualité intrinsèque.

La généalogie

L'étude sur la généalogie que l'on souhaite évoquer dans cette partie se limite à l'historique des modifications sur les données. Ainsi, dans les travaux de Keßler et al. (2011), les auteurs ont mis en œuvre une approche orientée vers les données rendant explicite toutes les modifications survenues sur une entité, à travers un vocabulaire de la provenance des données. Les contributions à OpenStreetMap sont organisées en des ensembles de modification qui contiennent des entités nouvelles, mises à jour et supprimées, éditées par un

utilisateur spécifique en une seule session. Le vocabulaire de provenance s'articule autour de 1ere édition d'une entité qui se caractérise par un ensemble des caractéristiques (tags, nouvelle géométrie, date d'édition, identifiant de *changeset*, numéro utilisateur), créant la 1ere version de l'entité. La 2nde édition reprend les informations de la 1ère édition (toutes les caractéristiques de la 1ere version de l'entité), la modifie ou la supprime et récrée une 2nde version de l'entité. Grâce à un graphe de restitution des états d'une entité, nous pouvant analyser l'historique d'Édition et de changement d'une 1ere contribution dans OSM. Chaque utilisateur a accès à l'instant présent, une collection de n modifications affectant n entités issus de la classe *Changesets*. Ce vocabulaire a permit aux auteurs de rendre explicite les informations de provenance implicites sur la lignée des entités dans OpenStreetMap et de les classer selon des modèles d'édition et de co-édition récurrents. Les auteurs proposent de formuler plu tard des fonctions de calcul de confiance et de réputation sur la base de ces reclassements d'instances. Les valeurs de confiance des entités peuvent être calculées sur la base de l'historique d'édition d'une entité tandis que les contributions d'un utilisateur et les modifications ultérieures des entités concernées par d'autres utilisateurs peuvent être utilisées pour mesurer la réputation de l'utilisateur.

Au fait, tous les états d'édition dune entité géographique dans OSM sont contenus dans un fichier historique [2]. La figure 3.4, illustre un extrait d'un fichier historique sur une édition d'un nœud. On remarque que ce nœud a été modifié 7 fois et donc admet 7 édition issues de plusieurs utilisateurs.

```
<node id="27128507" lat="50.0146754" lon="8.2429516" version="1" timestamp="2007-04-05T18:11:17Z"
  changeset="6911" uid="3609" user="seb"> <tag k="created_by" v="JOSM"/> </node>
<node id="27128507" lat="50.014677" lon="8.2429489" version="2" timestamp="2007-10-16T12:38:56Z"
  changeset="203178" uid="16643" user="Joh"/> </node>
<node id="27128507" lat="50.0144034" lon="8.2431315" version="7" timestamp="2013-03-16T23:41:50Z"
   changeset="15390280" uid="440308" user="spezialist"/> </node>
```

FIGURE 3.4 – Exemple d'extrait d'un fichier historique dans OSM ;Source : (Hashemi et Abbaspour, 2015)

Cohérence topologique

Parmi les travaux d'évaluation qui ont porté sur la cohérence topologique, Hashemi et Abbaspour (2015) ont proposé d'évaluer la qualité intrinsèque sur la cohérence topologique afin de détecter ds incohérences topologiques (de sorte à identifier l'objet erroné spatialement) à travers le concept de similarité spatial. Ce concept repose sur trois indicateurs à savoir la relation directionnelle (direction relative par rapport à son objet voisin), la relation topologique (disjoint, intersecte, contient equal,...etc) et une métrique de distance qualitative relationnelle basée sur le niveau de granularité (égal,proche, moyen éloigné) d'un objet géographique à ses voisins. Par la suite , les auteurs réalisent une multi-représentation d'un objet géographique par différents utilisateurs. Suivant ces multi-représentations, il apparait des changements sur les relations que cet objet entretient avec les autres objets. Évaluer la cohérence topologique revient à évaluer les relations qu'entretient un objet géographique avec ses voisins dans toutes ses différentes représentations. Ainsi, on mesure la similarité spatiale pour observer comment a évolué la cohérence topologique pour détecter une incohérence topologique d'un objet géographique.

2. http ://planet.openstreetmap.org/planet/full-history/

Complétude

Quant à la complétude, Barron et al. (2014) ont étudié l'historique des contributions et proposent une manière intrinsèque d'évaluer la complétude des données dans OSM. Sous l'hypothèse que l'édition du réseau routier s'effectue de manière chronologique d'abord par l'édition des autoroutes, puis les routes municipales et enfin les rues dans les zones résidentielles et toutes les autres routes telles que les chemins forestiers ou les petits sentiers, les auteurs identifient l'achèvement de l'édition d'une route à travers un état de stagnation de la longueur mensuelle d'une route tandis qu'au même moment une route de catégorie inférieure connaitrait une croissance du nombre de ses contributeurs. Cela pourrait marquer la fin de l'édition de la route en question pour passer à l'édition de celle de niveau inférieur. Ainsi on pourrait inventorier le nombre des routes complètes dans OSM sans être obliger de la comparer avec une base de données de référence. Les auteurs estiment que l'avantage ultime de cet indicateur peut être vu dans son indépendance d'un ensemble de données de référence qui le rend applicable pour n'importe quelle région du monde tout en reconnaissant que des déclarations absolues sur l'exhaustivité du réseau routier ne sont possibles qu'à l'aide d'un ensemble de données de référence de vérité de terrain.

Précision temporelle.
Sur la précision temporelle, Girres et Touya (2010) ont étudié la quantité d'objets mis à jour dans un jeu de données d'OSM en suivant l'évolution du nombre d'objets OSM mis à jour sur une période de trois mois (juin à octobre 2009) en France. Ils utilisent des statistiques pour observer les corrélations d'une part entre le nombre de contributeurs et la date moyenne de capture, et d'autre part la corrélation entre le nombre de contributeurs et la version moyenne de l'objet capturé. Les résultats montrent une augmentation globale d'environ 31,7% correspondants à plus de 260 000 objets saisis tout en précisant que les évolutions ont été principalement des ajouts de nouveaux objets plutôt que des mises à jour d'objets existants. Leur analyse a montré une nette augmentation linéaire de la date moyenne par rapport au nombre de contributeurs dans la zone : plus il y a de contributeurs, plus les objets sont récents. La même tendance est observée pour la version moyenne des objets capturés : plus il y a de contributeurs, plus les objets sont à jour.

3.3.6 Qualité basé sur le contexte spatial

L'évaluation de la qualité intrinsèque à travers le contexte spatial est un moyen de vérifier la véracité ou la fiabilité d'un information géographique. En effet, Jolivet et Olteanu-Raimond (2017) ont proposé un workflow sur laquelle les contributeurs d'une plateforme proposée par IGN France appelée Ripart, signalent des contributions (en guise d'erreurs, des remarques, des ajouts) dans le but de déclencher des mises à jour locales. La fiabilité de ces contributions doit d'abord être validée en évaluant leur cohérence avec le contexte spatial à l'aide des limites administratives et des bases de données topographiques ainsi qu'avec les attributs des ontologies prédéfinis. Les auteurs ont proposé des indicateurs dans ce sens.

Ainsi, toute contribution se voit d'abord évaluée par la cohérence de son emplacement géographique (la commune choisie à partir d'une liste de communes) avec la limite administrative qui l'inclut dans la base des limites administratives, suivi de la cohérence de l'emplacement avec le contexte spatial de sorte à intersecter le thème de la contribu-

tion avec le thème issu de la base topographique (exemple : le thème *bâti* du rapport ne doit pas se trouver dans le thème végétation ou hydrographie de la base topographique). Puis on vérifie si les commentaires faites sur la contribution voulant préciser le thème se concordent ou pas aux commentaires contenus dans un dictionnaire d'ontologie relatifs au thème de la contribution. Tous les résultats de ces différents indicateurs sont fusionnées pour déterminer la fiabilité de la contribution apportée par les citoyens capteurs ayant une certaine connaissance à une zone à cartographier.

Par ailleurs partant du constat que la base OSM intègre et rassemble d'objets ayant de niveau de détails variés (pistes cyclables, des routes maritimes, des zones bâties) dû au fait de la diversité des appareils de capture, Touya et Brando-Escobar (2013) donnent l'exemple de la coexistence entre un bâtiment détaillé et une parcelle afin d'illustrer que des incohérences provenant de niveau de détails provoqueraient des problèmes de lisibilité et parfois des incohérences au contexte spatiale. Selon une échelle de **Lickert**, les auteurs proposent 5 catégories de niveau de détails à savoir, **rue**, **ville**, **compté**, **région** et **pays**. Ils tentent de déterminer le niveau de détails des entités à travers un certain nombre d'indicateurs intrinsèques afin de les classer dans les 5 catégories. Une incohérence du contexte spatiale se décèle par la coexistence de deux entités classées dans deux catégories différentes.

3.3.7 Qualité basée sur l'apprentissage

A l'issue de ce qui a été dit, bien que la plupart des méthodes aient été utilisées pour évaluer la qualité extrinsèque, la qualité intrinsèque basée sur l'historique ou encore la qualité basée sur les contributeurs, il est à remarquer que peu de méthodes abordent le reste des indicateurs intrinsèques. La nécessité de fournir d'autres outils de calculs pour évaluer la qualité des données OSM n'est pas passée inaperçue (Yan et al., 2020). Dans ce sens, Senaratne et al. (2017), suggèrent que le Datamining (l'exploration de données) peut être utilisée comme une approche autonome, complètement indépendante des lois et des connaissances de la géographie, et indépendamment des approches sociales ou collectives pour évaluer la qualité de VGI. De ce fait, les auteurs proposent par exemple, la méthode PTV (Possibilistic thruth value)(De Tré et al. (2010)), utilisée pour évaluer l'incertitude de position des POI (points d'intérêts) basée uniquement sur la distribution possible. De même, les méthodes de détection des valeurs aberrantes (Ivanovic et al. (2019)), d'analyse de grappes (Jacobs et Mitchell (2020)), d'analyse de régression ou de corrélation peuvent être utilisées pour évaluer la qualité des données en décrivant purement et en apprenant sur des modèles de données.

Par ailleurs, on retrouve dans la littérature d'autres travaux qui ont conçu des approches permettant de prédire la qualité extrinsèque des données grâce à l'apprentissage automatique. Parmi ces approches, nous pouvons citer celles qui se basent sur des méthodes d'apprentissage en profondeur (Deep-Learning) utilisant plusieurs couches pour sélectionner certains descripteurs afin de reconstruire progressivement une sortie aussi similaire que possible à la couche de départ. Par exemple, Xu et al. (2017) ont utilisé un réseau d'auto-codage pour reconstruire la meilleure forme d'une variable grâce à la détection d'anomalies sur cette variable. Cette méthode utilise d'une part des données OSM et de référence, et d'autre part une image d'entrée pour détecter l'empreinte des bâtiments.

En outre, dans les travaux de Sturrock et al. (2018), les auteurs proposent une approche

basée sur l'apprentissage machine dans le cadre d'un programme de santé publique visant à pulvériser les habitations contre le risque d'apparition de malaria. Pour ce faire, et en guise de données d'entrainement, les auteurs étiquètent un échantillon de bâtiments dans OSM dont leur **type** est connu (résidentiel, commercial,..etc) avec deux labels à savoir **pulvérisable** ou **non-pulvérisable**. Les auteurs se base sur méthode d'apprentissage automatique dite *d'ensemble* (combinant plusieurs méthodes de classification) autour d'un certain nombre d'indicateurs intrinsèques caractéristiques aux bâtiments telles que la taille, la forme et la proximité des entités voisines, pour former un modèle permettant de prédire dans laquelle de ces deux classes se situerait chaque bâtiment. La taille et la forme caractéristiques d'un batiment résidentiel ou non-résidentiel,sont en fonction de superficie du toit de la structure, numérisée à partir d'images satellite et de la composition du toit, à savoir s'il s'agit d'un toit de chaume ou non. La proximité des bâtiments voisins à la route exprime également la nature des bâtiments (résidentiel ou non-résidentiel). En guise de validation du modèle proposé, les auteurs ont constitué un autre échantillon pris comme vérité terrain. Sur cette base de vérité, les résultats ont montré qu'au moins 86% des bâtiments (de la vérité terrain) ont été correctement classés par l'algorithme d'apprentissage.

Les études basées sur l'apprentissage automatique évoquées ci-dessus, ont le point le commun de combiner une approche extrinsèque et une approche intrinsèque pour évaluer la qualité d'un jeu de données d'OSM. Les études de ce groupe ont d'abord adopté une approche extrinsèque pour former des modèles appropriés à l'aide de métriques choisies, puis ont utilisé les modèles formés pour estimer, à l'aide d'une approche intrinsèque, la qualité des données OSM comme témoignent les travaux de (Yan et al., 2020). Il en est de même pour les travaux de Mohammadi et Malek (2015).

3.3.8 Conclusion sur l'état de l'art

L'évaluation de la qualité d'un jeu de données OSM se fait traditionnellement par comparaison avec un jeu de données d'une base de référence. Cette approche extrinsèque d'évaluation de la qualité a étudié par ordre d'importance la précision positionnelle, l'exhaustivité et la précision thématique l'accent étant principalement mis par ordre décroissant sur les routes OSM, les bâtiments, l'utilisation du sol et les points d'intérêts. En terme de précision de position, nous estimons que de nos jours les données OSM ont un écart de position très faible sur les données de type réseau, un peu moindre sur le surfacique (sur le bâti) et acceptable sur le ponctuel par rapport aux données faisant autorité. Néanmoins cette précision de position est entachée par une hétérogénéité spatiale globale. Quant à la précision thématique et sémantique d'OSM, nous observons qu'il manque beaucoup d'attributs et quand ils existent ils sont mal classés. Cela est dû par le fait que les contributeurs OSM annotent les objets géographiques en utilisant de manière inappropriée l'ontologie OSM (vocabulaire contrôlé), nuisant ainsi la qualité des données OSM. Pour pallier à ce problème d'annotation, nous partageons l'avis des chercheurs exprimant qu'il faudrait améliorer les plateformes de saisie d'OSM et de manière générale de VGI, afin d'éviter des erreurs d'annotation des attributs et/ou de classification.

Par contre il a été observé que les données OSM demeurent parfois même plus complètes (en terme d'objets géographiques) que celles des données d'autorité en urbain qu'en rural même si la taille des zones d'activité des contributeurs OSM est variable à l'échelle des régions ou des pays. En zone rurale, il existe moins de contributeurs d'où la fai-

blesse du nombre des contributions. Toutefois, la base OSM regorge beaucoup d'objets géographiques en urbain intéressant aux agences nationales souhaitant mettre à jour leurs bases de référence. Ces agences souhaitent profiter de l'avantage immédiat de capture des changements ou de nouveautés opérant sur l'espace géographique. Plu tard, le challenge réside dans la mutualisation ou l'interopérabilité des données VGI provenant de plusieurs sources VGI.

Dans un contexte de qualification des données OSM, la qualité OSM est évaluée à travers la crédibilité et la fiabilité du contributeur en se basant sur des approches sociales et du crowdsourcing. Les indicateurs sur les contributeurs issues de ces approches, demeurent à la fois difficile à les implémenter et subjectives de part leur définition même, et encore indirectes du fait qu'on infère la qualité de la contribution à partir de celle du contributeur.

Pour évaluer un jeu de données de manière intrinsèque et en se basant sur la dernière version mise à jour de la donnée, les chercheurs formulent des indicateurs qui se rapportent souvent la forme et la position des objets géographiques afin de déceler des imperfections témoignant d'une mauvaise qualité de saisie des objets géographiques. Parfois ils se basent sur le contexte ou la thématique étudiée afin de formuler des indicateurs détectant des aberrations, des erreurs de forme ou de position, etc. Ce qui permet de mesurer par la suite une qualité sur les données en question.

Quand nous nous appuyons sur une approche historique de données pour évaluer un jeu de données OSM, nous accédons au vocabulaire de provenance détaillant tous les états des entités géographiques ainsi que toutes ses modifications au cours temps. En se basant sur ce vocabulaire, les chercheurs sont en mesure de mettre en place des fonctions de calcul sur des indicateurs intrinsèques tels que la confiance et la réputation. Grâce à une étude de l'historique des données, les chercheurs parviennent à détecter des incohérences topologiques. De plus, suivant l'historique d'édition, nous pouvons évaluer la complétude des données dans une zone donnée par la localisation d'une transition marquant l'arrêt de contribution sur un type d'objet géographique donnée et le début de la contribution sur un autre type d'objet géographique d'une hiérarchie inférieure (exemple l'arrêt d'édition d'autoroute pour passer à l'édition d'une route), ainsi marquant la fin de l'édition de type d'objet en question. Enfin suivant le nombre d'objets mis à jours avec l'évolution du nombre des contributeurs, nous pouvons voir que les objets récemment mis à jour sont reliés avec une forte augmentation du nombre des contributeurs, ainsi exprimant la précision temporelle des objets géographiques en question. La plupart des études évaluant la qualité intrinsèque ont porté sur l'historique des données.

Une autre approche d'évaluation de la qualité peut se baser sur l'étude du contexte spatial pour vérifier la fiabilité d'une contribution en comparant son thème avec un autre thème au même niveau de détail. Quand il existe des différences des niveaux de détails, des incohérences au contexte spatiale peuvent apparaitre sur deux couches d'informations.

Une dernière approche consiste à évaluer la qualité des données OSM par apprentissage en faisant appel à des méthodes d'exploration de données visant à détecter des structures homogènes sur les données afin de déduire une mesure et/ou une classification de la qualité des objets géographiques d'OSM. Cette approche de l'apprentissage automatique peut parfois combiner une approche extrinsèque et une approche intrinsèque d'évaluation de la qualité des données.

Cependant, notre travail vise à fournir un estimateur intrinsèque de la qualité extrinsèque en utilisant uniquement la version actuelle de l'OSM et des données de référence. Notre approche consiste à trouver un lien statistique reliant les indicateurs de qualité extrinsèque aux indicateurs intrinsèques. Notre hypothèse de recherche repose sur le fait que les indicateurs intrinsèques portant sur les caractéristiques géométriques et positionnelles des objets géographiques permettent l'estimation des indicateurs extrinsèques de la qualité des données OSM. Les indicateurs intrinsèques sont formulés sur la base d'une analyse faite sur les conditions de saisies de données dans le but de définir des indicateurs intrinsèques, marqueurs d'une mauvaise/bonne saisie de données, capable d'estimer une mesure sur la qualité extrinsèque. Pour parvenir à concrétiser notre méthode, nous utilisons un certain nombre d'outils techniques que nous développons dans la section suivante.

3.4 Outils utilisés dans le processus de qualification de données géographiques

3.4.1 Appariement de données géographiques

3.4.1.1 Définition d'un processus d'appariement

Le processus d'appariement consiste en la mise en correspondance des objets géographiques homologues issus de deux bases de données. Ces objets représentent la même réalité sur le terrain et sont qualifiés de semblables par l'outil d'appariement. La notion de ressemblance entre les objets est évaluée à travers un certains nombre de critères en se basant sur une comparaison de lieu, de nature, de relation spatiale et de forme entre les objets à apparier. Toutefois la ressemblance demeure difficile à capter quand il existe des différences de simplifications géométriques du lieu (un bâtiment peut être représenté dans une base par un point et dans une autre par un objet surfacique) ou quand les deux bases de données ont des niveaux de détail différents. Cela diminue le degré de cohérence entre les objets à apparier. Dans le cas où les bases de données ont des niveaux de détails différents, on s'attend à une inclusion de la base moins détaillée dans la base la plus détaillée (Devogele, 1997; Mustière et Devogele, 2008). Si les deux bases de données ont des niveaux de détails et de contenu proches, on assiste à un chevauchement lors de l'appariement (Badard, 2000). Dans le pire des cas, où les deux bases de données sont sensiblement différentes en contenu et en niveau de détail, on obtient un chevauchement partiel (Raimond et Mustière, 2008).

Aussi, la variété des niveaux de détails entre les deux bases à apparier engendre une variété sur la notion de cardinalité des liens d'appariements. La cardinalité d'un lien d'appariement correspond au nombre d'objets mis en correspondance. En effet, pour le cas d'inclusion, on opère une cardinalité $1 : N$ (un objet de la base de données moins détaillée est mis en correspondance avec N objets dans la base la plus détaillé) tandis que pour le cas d'un chevauchement, la cardinalité $1 : 1$ (un objet d'une base de données est mis en correspondance avec un seul objet dans l'autre base de données) est de mise. Enfin, pour un chevauchement partiel, on s'attend à une cardinalité $N : M$ (N objets d'une base de données sont mis en correspondance avec M objets dans l'autre base de données).

Cette étape d'appariement est nécessaire pour mener à bien plusieurs traitements sur les données géographiques par exemple l'évaluation de la qualité, le recalage des données géographiques, la mise à jour des bases de données et enfin l'intégration d'une base de

données dans une autre base de caractéristiques différentes.

Ainsi, le processus d'appariement sert à recaler des données géographiques dans un référentiel géographique, le but étant d'assurer l'Interopérabilité entre deux jeux de données différents (quand un des jeux contient des données thématiques par exemple). Le principe consiste à superposer deux jeux de données représentant la même réalité. Après appariement, on réalise une transformation géométrique sur l'un des deux objets appariés pour le superposer à l'autre. On retrouve dans la littérature des travaux qui ont procédé à apparier des points de contrôle de type vecteur avec des images dans le cadre d'un processus de géoreferencement des orthophotos (Saalfeld, 1988). Les travaux de Davis et Fonseca (2007) s'intéressent aux recalages des *adresses* sur une base d'adresses ayant une référence spatiale en utilisant des mesures de similarité entre les chaînes de caractères telles que la distance de Levenshtein (Levenshtein, 1966). Le concept d'adresses dans cette approche correspond à la fois aux adresses postales et à d'autres informations qui font référence à un lieu telles que le nom du bâtiment, le code postal, le code téléphoniques, etc. (Davis et Fonseca, 2007)

Dans le cadre de mises à jour des bases de données de référence, on fait appel au processus d'appariement pour détecter l'évolution entre deux actualités dans deux bases de données (l'une ancienne l'autre étant la base dérivée ou récente) se matérialisant au sens d'une cardinalité. Si l'appariement détecte une cardinalité 1 : 0, il s'agit d'une suppression de l'objet existant dans la base de données récente tandis pour une cardinalité 1 : 1, il s'agit d'une scission, c'est à dire dire qu'il n'y a pas de mise à jour à faire, les deux objets étant représentés dans les deux bases. La cardinalité $N : 1$ signifie qu'il y eu une fusion des objets de la base ancienne dans la base récente. Si la cardinalité est $N : M$, on assiste à des modifications dans les deux bases et enfin une cardinalité 0 : 1, renseigne l'apparition d'un nouveau objet dans la base récente. Toutefois, dans ce contexte de mises à jour, les deux bases de données ont le même niveau de détails et d'hétérogénéité est réduite au niveau de la représentation des objets de base, de la sémantique et de la géométrie. On précise que les objets n'ayant pas subi des changements ont gardé la même position géographique (Olteanu-Raimond, 2008; Ivanovic et al., 2019).

Enfin, le processus d'appariement est utilisé dans le but d'intégrer des bases de données hétérogènes afin d'unifier la sémantique des deux bases de données, d'éliminer les objets redondants et enfin de regrouper les objets similaires. Cet appariement entre bases hétérogènes peut passer par l'établissement d'une correspondance entre les schémas de données des deux bases. La correspondance peut se faire en comparant les classes d'objets des deux bases et en dégageant des ressemblances entre ces classes à travers une matrice de confusion (Olteanu-Raimond, 2008).

3.4.1.2 Classification des approches d'appariement des données

Pour tenter d'apporter une classification des approches d'appariement, nous nous appuyons d'une part sur les travaux de Xavier et al. (2016) qui ont réalisé une synthèse d'état de l'art des trente dernières années (jusqu'en 2016) sur la question d'appariement géospatiale tout en proposant des taxonomies autour duquel les auteurs classifient les approches d'appariement et d'autre part, les travaux de Costes et Perret (2019) qui analysent et classifient les approches autour de quatre points ou propriétés. Bien d'autres classifications ont été élaborés dans le passé par d'autres travaux de recherche, nous estimons que ces deux références bibliographiques que nous nous basons, semblent être à la fois

fédératrices et exhaustives sur l'ensemble des approches d'appariement mais aussi pertinentes pour notre processus d'appariement des données. Toutefois, nous constatons qu'il existe des chevauchements sur les deux classifications proposées par les deux références bibliographiques. C'est pourquoi, nous souhaitons noter les points de concordance afin de rallier les deux classifications. Après cela, nous relevons dans la littérature en guise d'exemple, un groupes d'approches d'appariement que l'on reparti dans les différents niveaux de classification.

Parmi les critères de classification proposée par Costes et Perret (2019), une première propriété d'une approche d'appariement concerne la prise en compte de l'imperfection lors de l'appariement à travers un traitement de cardinalité des liens d'appariement. Dans les travaux de Xavier et al. (2016), ce critère est appelé *cas de correspondance*. En effet, une imperfection s'exprime en une imprécision liée à la localisation des objets géographiques (un objet de grande taille représenté par un point), en une incertitude représentant le doute que l'on a sur la validité d'une information et enfin par l'existence d'une incomplétude issue de la différence du niveau de détails en deux jeux données ou/et d'une absence d'information sur un champ non rempli (dans un jeu de données plus détaillé, on a plus d'objets géographiques que dans le jeu le moins détaillé). Ainsi, l'incomplétude spatiale (c'est-à-dire l'absence d'objets géographiques) est traitée en prévoyant une cardinalité *1 :0* tandis que l'imprécision et l'incertitude sont prises en compte par le traitement de la précision spatiale (Olteanu-Raimond, 2008). Quand l'incomplétude touche aussi les attributs (information manquante), dans ce cas, le lien *1 :0* ne pourra plus gérer l'incomplétude d'une information thématique, temporelle ou sémantique. Hormis ce cas de cardinalité nulle prévu par défaut pour gérer l'incomplétude spatiale, l'appariement des données prévoit une obtenir une cardinalité *1 :1* ou *1 :n* ou enfin *n :m*.

L'appariement de données basé sur une cardinalité 1 : 1 a été le cas considéré au tout début des études d'appariement. C'est le cas des travaux de Beeri et al. (2004) et plus tard ceux de Safra et al. (2013), Song et al. (2011) et Fan et al. (2014). La cardinalité 1 : 1 n'est plus applicable lorsqu'un objet géographique est susceptible d'exister en plusieurs parties dans un autre jeu de données $(1 : n)$ (Raimond et Mustière, 2008; Wenjing et al., 2008), ou lorsque plusieurs objets géographiques d'un jeu de données semblent être homologues à plusieurs autres dans l'autre jeu de données $(n : m)$ (Zhang et al., 2014; Tong et al., 2014).

Une deuxième propriété selon Costes et Perret (2019), concerne la définition de la ressemblance entre deux objets géographiques sous l'hypothèse de ressemblance à travers des critères d'appariement. Cette propriété est évoquée dans Xavier et al. (2016) comme critère de classification en distinguant les mesures géométriques, les mesures topologiques, les mesures attributaires, les mesures sémantiques et enfin les mesures sur le contexte.

Les mesures géométriques se réfèrent à la position, la forme et le recouvrement des objets géographiques en utilisant des distances mesurant des écarts de position telles que la distance euclidienne, la distance de Hausdorff, la distance de Fréchet (Devogele, 2002; Mascret et al., 2006), ou la distance surfacique (Alt et Godau, 1995; Bel Hadj Ali, 2001a), des écarts de forme telles que la distance radiale et la distance angulaire (Bel Hadj Ali, 2001a).

Les mesures topologiques portent sur les relations topologiques notamment les travaux de Franzosa (1995) qui ont proposé un modèle d'équivalence topologique. Nous pouvons également citer les travaux de Raimond et Mustière (2008) ont proposé le critère de voi-

sinage comme mesure topologique. Ce critère est basé sur l'hypothèse que si deux arêtes sont similaires, alors leurs voisins devraient également être similaires.

Les mesures attributaires font référence à l'évaluation des propriétés non géométriques d'un objet géographique et se basent sur soit une distance numérique (à valeur 1 si deux termes sont identiques et 0 sinon) (Deza et Deza (2006)), distance sur liste de domaines fondée une table de correspondance des domaines des termes à comparer (cette distance varie entre 0 et 1 selon que les deux termes appartiennent ou non à des domaines proches) (Cobb et al., 1998) et enfin une distance comparant les caractères composant les deux termes relevant des écarts d'attributs telles que distance de Hamming ou la distance de Levenshtein (Hamming, 1950; Levenshtein, 1966).

Les mesures sémantiques tentent d'établir des distances sur la natures des objets géographiques à comparer. La sémantique peut être étudiée à travers des taxonomies classifiant certains attributs (exemple *nature*) en se basant notamment sur la distance de Wu et Palmer (1994). Enfin, Les mesures de contexte nous permettent d'évaluer la similarité entre des entités géographiques et le contexte géographique de la zone dans laquelle les entités se situent. Selon Samal et al. (2004), le contexte géographique fait référence aux relations spatiales entre les objets dans une zone, notamment les relations entre un objet et un ensemble limité de points de repère. Par exemple, Zhang et al. (2014) ont suggéré d'utiliser la triangulation de Delaunay afin de définir une distance sur le voisinage des objets géographiques en admettant qu'il existe une influence continue entre les objets les plus proches. Xavier et al. (2016) affirment que les mesures de contexte s'avèrent utile lorsque lorsqu'il y a peu d'informations sur les ensembles de données évalués, ou lorsqu'il existe un grand décalage entre eux. Dans ces cas-là, le contexte peut aider à réduire l'incertitude lors de la recherche des entités correspondantes.

Une troisième propriété d'une approche d'appariement porte sur son nombre de critères d'appariement. Plus le nombre de critères est élevé, plus compliqué est le processus et plus l'algorithme est compliqué à calibrer. Cette disparité est susceptible de conduire à des résultats différents en fonction des différents ensembles des mesures de similarité selon qu'ils soient de même nature ou différente (géométrique, topologique, sémantique, etc.) (Costes et Perret, 2019). Cette propriété est aussi mentionnée brièvement dans la classification faite dans Xavier et al. (2016) en distinguant si l'approche d'appariement utilise une seule mesure de similarité ou plusieurs.

Une dernière propriété proposée par Costes et Perret (2019) comme un critère de classification, concerne la combinaison de mesures de similarité pour la prise de décision par le processus d'appariement. Il s'agit de décider si les liens d'appariement sont établis séquentiellement ou simultanément. Il s'agit aussi de définir une manière de combiner les critères d'appariement et de définir à partir de quels seuils il faut établir un lien d'appariement ou au contraire retourner une réponse d'indécision. Cette propriété est bien détaillée dans Xavier et al. (2016) en analysant la méthode de combinaison des mesures de similarité ou critères d'appariement. Hormis le cas où les mesures de similarité (une seule mesure ou plusieurs) sont étudiées séquentiellement nécessitant à aucune combinaison (Beeri et al., 2005), les auteurs distinguent le cas où les critères d'appariement sont combinées en se basant sur des scores normalisés (Pendyala, 2002), le cas où est effectuée une combinaison des critères avec pondération (Zhang et Meng, 2007), le cas d'une combinaison basée sur la théorie de probabilité (Tong et al., 2009), le cas d'une combinaison basée sur un processus

optimisé (Li et Goodchild, 2012) ou le cas d'une combinaison fondée sur la théorie des croyances (Olteanu-Raimond et al., 2015).

Dans ce qui suit, nous passerons en revue les approches d'appariement qui nous semblent intéressants pour notre problématique de manière à les décrire selon les propriétés évoquées ci-dessus.

Parmi les approches, on dénote l'approche proposée par Walter et Fritsch (1999). Cette approche est destinée à apparier des réseaux ayant le même niveau de détail en se basant sur des filtrages statistiques et l'information mutuelle partagée par deux arcs à apparier. L'information mutuelle correspond à la probabilité totale de vraisemblance des deux arcs (qui détermine la décision ou non d'apparier ces arcs) à travers quatre mesures de similarité de type géométrique à savoir la localisation, longueur du réseau, la forme et l'angle. Ainsi le processus d'appariement prévoit une cardinalité $n:m$ mais en revanche nécessite un prétraitement sur le recalage de réseau. Lors de la combinaison, des seuils de sélection des candidats sont déterminés de manière empirique pour chaque mesure de similarité à travers un jeu de données de vérification.

Une autre approche a été proposée par Mustière et Devogele (2008) afin d'apparier deux jeux de données de type routier ayant de niveaux de détails différents en se basant sur des mesures de similarités géométriques et topologiques. Par rapport à l'approche précédente, cette approche réalise d'abord un appariement des nœuds pour finir par un appariement d'arcs. Le processus sélectionne des nœuds candidats du réseau le plus détaillé vers le réseau le moins détaillé en utilisant un seuil sur la distance euclidienne et inversement pour les arcs candidats à l'aide de la demi-distance de Hausdorff. Cela donne des nœuds et arcs pré-appariés. On retient parmi plusieurs pré-appariement des arcs, l'appariement des deux arcs ayant la plus faible distance et dont les nœuds respectifs sont déjà appariés tout en vérifiant que les arcs pré-appariés sont orientés de la même façon que les nœuds pré-appariés. Ainsi ce processus d'appariement s'effectue de manière séquentielle en prévoyant une cardinalité de type $1:n$ tandis que l'imprécision est gérée à travers des seuils pour la sélection des candidats. Ces seuils sont déterminés à partir d'une étude de sensibilité sur les seuils pour les mesures de similarité permettant une combinaison séquentielle adéquate.

Si les deux approches précédentes ont été proposées afin d'apparier deux réseaux, on relève dans la littérature l'approche voulant apparier des données surfaciques issues des travaux de Atef (2001). Le processus commence par une intersection de deux jeux de données. Puis la distance surfacique exprimant la différence symétrique des surfaces des deux objets surfaciques est calculée comme un critère d'appariement. Selon un seuil fixé sur la distance surfacique, le processus supprime les liens parasites ou incorrects. Puis s'ensuit une étape de calcul d'une matrice d'association permettant de distinguer deux types de liens d'appariement : les liens de cardinalité 1 : 1, et les liens de cardinalité multiple $N:M$. Enfin à ce processus peut être ajoutée une étape de raffinement des liens multiples en recherchant le meilleur groupe minimisant la distance surfacique. Ainsi, dans cette approche, l'imprécision est gérée à l'aide du seuil fixé pour l'élimination des liens parasites tout en adoptant une cardinalité $N:M$. L'appariement se fait de manière séquentielle et progressive.

Afin d'apparier plusieurs jeux de données, les travaux de Samal et al. (2004) se sont fondés sur une méthodologie rassemblant des données raster et vecteur mais aussi des

critères géométriques, sémantiques et topologiques. En premier lieu, l'approche considère l'appariement des objets en se basant sur des mesures sur les attributs ou la géométrie de manière à combiner l'ensemble de ces mesures et à retenir le couple d'objets ayant la valeur maximale de la somme pondérée des mesures issues de chacun des critères. Un critère topologique est ensuite calculé à partir d'une structure de graphe reliant les objets à apparier suivant des critères de proximité et d'orientation. L'appariement est déterminé par une combinaison des critères géométriques et du critère topologique de sorte à mener une étude de sensibilité sur les critères d'appariement ainsi que sur les poids accordés entre les critères. La cardinalité de type 1 : 1 est appliquée afin de créer de retrouver un lien d'appariement de cardinalité de manière progressive et itérative. L'imprécision est prise en compte par la combinaison de plusieurs critères d'appariement et par le processus itératif voulant apparier un objet avec celui ayant le maximum de ressemblance. Le processus d'appariement est simultanée : les critères d'appariement sont pris en compte simultanément mais les liens sont créés séquentiellement. Pour combiner ces critères, les auteurs ont du mené une étude de sensibilité sur les seuils de sélection des candidats.

Dans le but de gérer mieux l'imperfection dans sa globalité et d'expliciter les connaissances et l'ignorance quand ces connaissances n'existent pas, une nouvelle approche capable de fusionner plusieurs critères d'appariement a été proposée dans les travaux de Olteanu-Raimond et al. (2015). Les connaissances sont rendues explicites à travers des fonctions de croyance établies sur les critères d'appariement exprimant nos connaissances sur la réalité. L'approche multi-critère de la comparaison des données consiste à définir différents critères de comparaison. Pour chaque critère d'appariement, trois fonctions de croyance sont définies pour chacune des trois hypothèses, à savoir l'hypothèse *le candidat est l'objet homologue* ($appCi$), l'hypothèse *le candidat n'est pas l'objet homologue* ($-appCi$) et l'hypothèse d'ignorance disant *je ne sais si le candidat est l'objet homologue* (Θ). Cela exprime bien la croyance que l'on accorde à un candidat à travers ces hypothèses et se matérialise par ce que l'on appelle une masse de croyance. Ainsi, comme illustré sur la figure 3.5, le processus d'appariement commence par la sélection des candidats dans un disque de rayon fixé autour de l'objet à apparier. S'ensuit l'étape d'attribution des masses de croyances pour chaque candidat par critère et selon les trois hypothèses. Puis pour chaque candidat, le processus fusionne les critères à travers la fusion des masses de croyance. Ensuite, une étape de fusion des candidats permet de classer les candidats par ordre de croyance à l'objet à apparier par ordre de probabilité, cette probabilité étant déterminée par les masses de croyance. Enfin, l'étape finale consiste à prendre une décision finale selon trois cas : $Apparié$ ou $Non - Apparié$ ou $Indécis$. Pour le 1er cas, il existe un candidat qui se distingue nettement des autres candidats et donc l'algorithme choisit un candidat pour objet homologue tandis que dans le second cas, aucun des candidats ne présente suffisamment des similitudes avec l'objet à apparier et donc l'algorithme décide de ne pas apparier l'objet en question. Cela veut dire que le vrai objet homologue n'existe pas parmi les candidats. Dans le dernier cas l'indécision s'impose. Cela est du au fait que deux candidats sont soutenus par des critères différents de la même manière ou la différence de probabilité pignistique entre deux candidats classés par ordre de préférence est inférieure à un seuil de séparation pour les deux candidats. Ainsi, l'approche proposée dans Olteanu-Raimond et al. (2015), se base sur la théorie des fonctions de croyance et a été testée sur des données ponctuelles et linéaires prévoyant une cardinalité de type *1 :n* pour l'appariement d'objets linéaires. La combinaison est assurée par l'application du principe de la théorie des croyances.

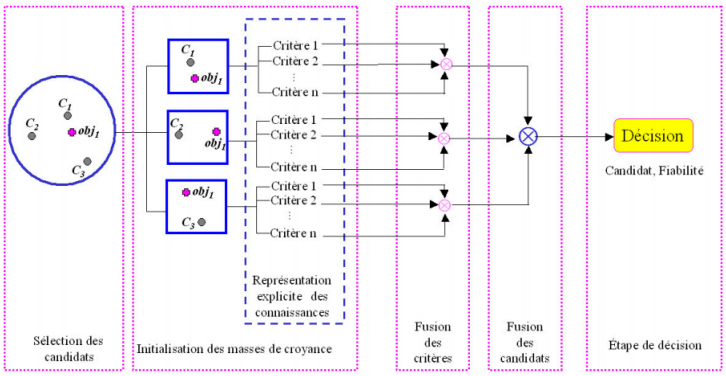


FIGURE 3.5 – Processus d'appariement basé sur la théorie de croyance ;Source : (Olteanu-Raimond, 2008)

Dans une perspective d'intégrer des données de différentes sources portant sur une même zone, un processus qualifié de confluence permet de créer une nouvelle base de données à partir de l'intégration de deux ou plusieurs bases de données afin d'obtenir des données de bonne précision, d'éliminer les redondances et de réconcilier les données en conflit. Ce processus passe par une étape d'appariement des données issues de plusieurs bases de données. Ainsi, les travaux de Tong et al. (2009) ont proposé une approche (par extension des travaux Beeri et al. (2004)) intégrant plusieurs mesures en se basant sur la théorie de probabilité. Leur méthode consiste à calculer une valeur de probabilité indiquant la ressemblance pour chaque paire d'objets potentiellement homologues par critère d'appariement, pouvant s'appliquer à des données ponctuelles, linéaires et surfaciques. La probabilité totale pour un couple d'objets homologues, est obtenue par la moyenne pondérée de l'ensemble des mesures issues des critères d'appariement. Des critères géométriques (distance euclidienne entre deux objets), attributaires et topologiques (la différence d'angles entre les diagonales des MBR (le minimum rectangle englobant) des deux objets) ont été utilisés dans cette approche dont l'incertitude est gérée à travers la combinaison des plusieurs critères d'appariement. Les données ponctuelles sont d'abord appariés suivi des données linéaires ou surfaciques. La combinaison est assurée par l'application du principe de la théorie de probabilité ainsi qu'une méthode de combinant simultanément les mesures de similarité pondérés dont leur poids ont été déterminées à travers une analyse de sensibilité sur les poids accordés aux critères.

Fan et al. (2016) ont introduit une nouvelle approche appariant deux groupes de blocs de bâtiments afin d'apparier au final un réseau routier issu d'OSM avec un réseau issu d'une source d'autorité. Les auteurs ont présenté une approche pour évaluer la qualité des données de bâtiments 2D disponibles dans OpenStreetMap pour la ville de Munich (Allemagne). Leur approche de mise en correspondance des entités surfaciques du bâti repose sur un recouvrement d'aire bidirectionnel, une technique capable d'identifier des objets homologues avec une cardinalité de type *n :m*. Sur chacun des deux réseaux à apparier les blocs urbains sont constitués par un processus de planarisation du réseau routier. Les blocs sont les faces de ce réseau planaire. Puis les arêtes (arcs du réseau) de chaque bloc urbain sont numérotés. Ensuite chaque arc du réseau (du jeu OSM et du jeu d'autorité) est lui aussi relié au bloc urbain de l'arête qui lui est la plus proche. Après superposition, les blocs urbains issus de deux jeux de données sont appariés si leur taux de chevauchement

supérieur à 50% et si une relation de cardinalité de type *1 :1* a été identifiée. Sur un couple de blocs urbains appariés, on recherche les arrêtes correspondantes. Si deux arcs des deux réseaux à apparier sont reliés aux arrêtes correspondantes, ces arcs deviennent à leur tour appariés. De cette façon, l'incertitude est réduite du fait que l'appariement tourne autour du bloc urbain et que les arcs à apparier sont ceux avoisinants au bloc urbain.

Cette approche utilise comme critère d'appariement d'une part, la distance euclidienne entre arcs (ou entre arêtes) pour rattacher à la fois un arc d'un réseau à une arête du bloc urbain et pour la recherche des arêtes correspondantes, et d'autre part une mesure de chevauchement entre les blocs urbains à apparier. Le processus d'appariement se fait de manière séquentielle. L'appariement des blocs urbains se caractérise la plupart du temps par une cardinalité 1 : 1. Toutefois, on peut observer une cardinalité prévoyant des relations variables du 1 : 1 au 1 : N ou N : 1 selon que l'un de deux jeu de données a été mis à jour ou selon que les deux jeu de données ont été mis à jour mais à différentes périodes ou à des complétudes différentes. Une analyse de sensibilité sur les seuils de sélection des candidats a été menée pour l'ensemble des mesures de similarité utilisées dans cette approche.

En dernier, on relève dans la littérature l'approche proposée par Costes et Perret (2019). Cette approche se base sur l'élaboration du modèle de Markov caché (MMC). Pour apparier un réseau G_1 avec un autre réseau G_2, on considère les arcs de G_1 à apparier comme étant les observations tandis les arcs de G_2 constituent les états. Chaque état admet une probabilité de correspondance avec une observation. Cette probabilité est à la base indépendante des autres probabilités que les autres états ont avec la même observation. Cette probabilité de base pour chaque état, est calculée selon des critères géométriques. Partant d'une séquence d'observations (une succession d'arcs de G_1), l'algorithme de MMC a pour but de trouver la séquence d'états (cachés) la plus probable. Cela se résout avec l'algorithme de Viterbi. On commence par apparier un 1er arc de G_1 avec un arc de G_2. Les probabilités d'appariement sont recalculés en fonction de cet état connu et avec l'aide des critères topologiques, et ce de manière itérative de sorte que l'appariement prochain se base sur l'appariement courant ayant eu lieu. Des correspondances multiples peuvent être observées.

Si l'on tente d'analyser la mise en œuvre des approches ci-dessus, on remarque que la plupart des algorithmes implémentant les approches sont testés sur des données linéaires (réseau routier). Parmi ceux là, quelques-uns traitent le réseau routier avec différents niveaux de détails (Tong et al., 2014; Mustière et Devogele, 2008) ou sont adaptés à des objets aux géométries différentes (Walter et Fritsch, 1999). Certaines approches ont été proposées pour apparier des objets ponctuels (Beeri et al., 2004; Samal et al., 2004) ;(Olteanu-Raimond et al., 2015). Plus tard des approches appariant des entités surfaciques apparaissent (Fan et al., 2016; Ruiz-Lendínez et al., 2016; Kim et al., 2010).

La plupart de ces approches utilisent des mesures de similarité de type géométrique faisant souvent appel respectivement par ordre d'importance, à la distance euclidienne (Beeri et al., 2004; Mustière et Devogele, 2008) aux distances sur la forme (Walter et Fritsch, 1999), la distance de Hausdorff (Huh et al., 2011; Mustière et Devogele, 2008) et mesure sur le recouvrement d'objets. Quelques fois, les approches ont appariés des données grâce à des mesures de similarité de type topologiques (Tong et al., 2009; Samal et al., 2004).

Nous avons également observé que la plupart de ces approches tentent de déterminer

des seuils de décision sur les critères d'appariement (Zhang et al., 2005) en menant une analyse de sensibilité sur seuils (Samal et al., 2004; Beeri et al., 2004; Fan et al., 2014; Ruiz-Lendínez et al., 2016; Mustière et Devogele, 2008) relatifs aux critères, et assez des fois les approches recherchent des seuils de pondération basés toujours sur une étude de sensibilité sur les poids (Tong et al., 2009) accordés aux critères.

Enfin, les approches détectent le plus souvent des liens de cardinalité de type 1 : n dans un cadre de comparaison des deux jeux de données ayant différents niveaux de détails ou dans le cas des mises à jour des bases de données ou encore dans le cas d'intégration des données de différentes sources (Walter et Fritsch, 1999; Mustière et Devogele, 2008; Wenjing et al., 2008). Néanmoins, le cas des approches opérant avec une cardinalité de type 1 : 1 semble être aussi répandus (Ai et al., 2013; Li et Goodchild, 2012).

3.4.1.3 Métriques pour apparier des objets géographiques surfaciques

Dans le cadre d'un processus d'appariement des données, il est nécessaire de se munir des métriques sur lesquelles, la recherche des objets homologues devrait se baser. Comme déjà mentionné dans 3.4.1.2, on peut identifier des métriques basées sur la géométrie (forme et position), la thématique, la sémantique et la topologie. Certaines de ces métriques sont définies pour tous les types d'objets géographiques (*ponctuel, linéaire et surfacique*) tandis que d'autres sont applicables qu'à un seul type d'objets géographiques. Vu notre besoin annoncé dans le chapitre **??**, nous nous intéressons qu'aux métriques permettant de mesurer des ressemblances entre des objets surfaciques.

Pour appréhender une qualité basée pleinement sur les caractéristiques des entités surfaciques, Atef (2001) suggère que le contrôle de la qualité des entités surfaciques ne devrait pas se baser uniquement sur les écarts de position par rapport aux données de contrôle ou à la base de référence, mais aussi sur les écarts de forme. Ainsi, afin de mieux refléter la qualité spatiale issue d'un appariement de données surfaciques, il est important de choisir des mesures qui se basent sur la position et sur la forme des entités surfaciques évaluant la ressemblance de deux objets à apparier. L'écart entre deux mesures relatives à deux objets constituera une métrique ou une distance [3]. Dans ce qui suit, nous évoquons les métriques utilisées dans le cadre de cette thèse lors de l'appariement des données.

Afin d'apparier des bases de données surfaciques, les approches proposées s'appuient sur des mesures basées sur la forme et sur la position des objets géographiques. En guise de mesure sur la forme, on se sert tout d'abord de la distance polygonale basée sur la fonction polygonale.

Pour cela, sur un polygone, on mesure, pour un réel $s \in [0, 1]$ associé à chaque sommet du polygone, la distance entre le centre de masse du polygone et le point (sommet) d'abscisse curviligne sP en parcourant le polygone dans le sens trigonométrique direct avec P étant le périmètre du polygone. L'abscisse curviligne est normalisée pour que la fonction soit définie sur $[0, 1]$) (Bel Hadj Ali, 1997). La fonction polygonale, pour un point d'origine des mesures (point de départ), se définit par Bel Hadj Ali (2001a) :

$$\mathbf{S} : [0,1] \rightarrow \mathbb{R}^+$$

3. Le plus souvent, on pourra prouver que cette métrique a bien les propriétés d'une distance mathématique.

$s \rightarrow S(t) = \sqrt{(x_c - x(t))^2 + (y_c - x(t))^2}$; x_c et y_c sont les

coordonnées du centre de masse du polygone, $x(t)$, $y(t)$ étant les coordonnées du point du contour d'abscisse curviligne $t = sP$.

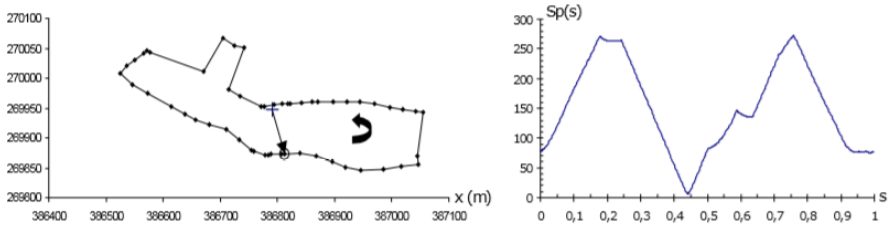


FIGURE 3.6 – Signature polygonale ; Source : Bel Hadj Ali (2001a)

D'après la formule 3.6, nous remarquons que la fonction polygonale ne dépend que des distances. Et sachant que les distances sont insensibles à la rotation et à la translation, on peut dire que la signature polygonale se caractérise par une invariance à la translation et à la rotation mais dépend du point de départ. Ainsi, pour un polygone donné, il peut y avoir deux signatures polygonales issues de deux points de départ différents et ayant un déphasage égal à la distance qui sépare les deux points de départ, normalisé par le périmètre du polygone. Par contre, la signature polygonale demeure sensible à l'homothétie qui se traduit par une multiplication par un facteur k à la signature polygonale d'un polygone ayant subi une transformation d'homothétie.

La distance radiale entre deux polygones correspond à l'intégrale de la différence de leurs deux fonctions polygonales. Cependant, le calcul d'une distance entre deux signatures polygonales varierait à chaque fois que les points d'origines changent. Afin de remédier au problème du point d'origine, nous choisissons le décalage qui minimise l'intégrale de la différence des deux fonctions polygonales.

Soient A et B deux polygones, supposons que le point origine des mesures est décalé d'une quantité τ le long du contour du polygone B, la distance $d_r(A, B)$ se traduit par :

$\forall\ (A, B) \in \mathcal{F}^2$,

$$d_r(A, B) = \left(\min_{\tau \in [0,1]} \int_0^1 \left(S_A(t) - S_B(t + \tau) \right)^2 dt \right)^{\frac{1}{2}}$$

où $S_A : [0, 1] \rightarrow \mathbb{R}^+$ (resp. S_B) désigne la signature polygonale de A (resp. B).

La seconde distance relative à la forme que nous avons utilisée dans notre processus d'appariement est celle basée sur la fonction angulaire nommée *distance angulaire*.

Pour une valeur $x \in [0, 1]$, la fonction angulaire est l'angle entre la tangente au polygone au point d'abscisse curviligne px et l'axe horizontal, avec p étant le périmètre du polygone.

La fonction angulaire est invariante à la translation à la rotation et à l'homothétie mais dépend du point d'origine. Deux fonctions angulaires d'un même polygone déterminées à partir de deux points d'origines différents sont égales à un déphasage en x près.

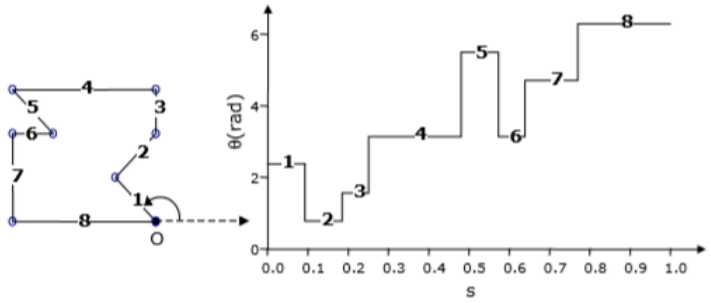


FIGURE 3.7 – fonction angulaire ;Source : Bel Hadj Ali (2001a)

La distance angulaire entre deux polygones se définit comme étant l'intégrale des différences de leurs deux fonctions angulaires. Afin d'éliminer le déphasage, nous choisissons le décalage qui minimise l'intégrale. On précise tout de même que la fonction angulaire ne s'applique pas à des polygones à trous ou polygones complexes. La distance angulaire s'écrit comme suit :

$\forall\ (A, B) \in \mathcal{F}^2$,

$$d_a(A,B) = \Big(\min_{\tau\in[0,1]} \int_0^1 \big(\theta_A(t) - \theta_B(t+\tau)\big)^2 dt\Big)^{\frac{1}{2}}$$

où $\theta_A : [0,1] \to \mathbb{R}$ (resp. θ_B) désigne la signature angulaire de A (resp. B).

Pour les distances basées sur la position, notre processus d'appariement utilise la distance surface et la distance de Hausdorff.

La distance surfacique de deux entités surfaciques correspond au ratio de l'aire de la différence symétrique des deux entités et de l'aire de leur union. Si les deux entités sont complètement disjointes, leur distance surfacique vaut 1. Par contre, si elles sont parfaitement égales (superposées), leur distance surfacique vaut 0.

Quant à la distance de Hausdorff, elle s'établit entre deux entités surfaciques de manière à retrouver la distance maximale d'éloignement entre les deux surfaces en se basant sur l'intérieur des entités surfaciques et non le contour comme illustré dans la figure 3.8. La distance de Hausdorff définie tout d'abord dans Hausdorff (1937) comme une distance euclidienne (Abbas, 1994), et étendue dans Atef (2001) pour les entités surfaciques, s'énonce dans ce cadre de thèse comme suit :

$$dh(A,B) = \max\Big(\max_{x\in A} d(x,B), \max_{y\in B} d(A,y)\Big)$$

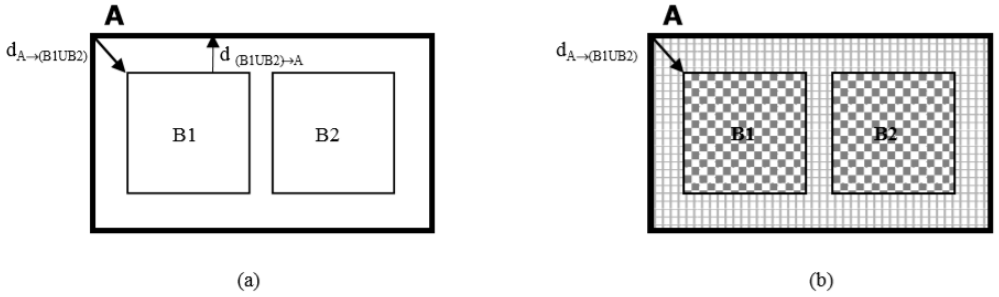


FIGURE 3.8 – Exemple distance de Hausdorff : Source : Atef (2001). En se basant sur le contour (a), le calcul de la distance de Hausdorff renvoie la valeur maximale des deux quantités $dA \to (B1 \cup B2)$ et $d(B1 \cup B2) \to (A)$ tandis si l'on se base de l'intérieur des entités surfaciques, la quantité $d(B1 \cup B2) \to (A)$ s'annule et la distance Hausdorff est égale à $dA \to (B1 \cup B2)$.

3.4.2 Méthodes d'apprentissage statistique

3.4.2.1 Définition du problème d'apprentissage

Dans un contexte d'évaluation de la qualité de données géographiques, les recherches récentes font appel aux méthodes d'apprentissage statistique afin de remédier au problème de l'absence d'une base de référence avec laquelle on mesure la qualité extrinsèque. L'objectif principal étant de prédire une qualité extrinsèque sur un jeu de données non évaluées auparavant, l'apprentissage statistique permet de formaliser un processus capable d'apprendre à partir des données, l'estimation d'une mesure exprimant une mesure de qualité, telle que la précision spatiale.

L'apprentissage statistique comporte généralement deux phases. La première phase consiste à estimer un modèle à partir de données, appelées observations, qui sont disponibles et en nombre fini, lors de la phase de conception du système. L'estimation du modèle consiste à résoudre une tâche pratique, telle que traduire un discours, estimer une densité de probabilité, reconnaître la présence d'un chat dans une photographie, classer des images de tumeurs, identifier les facteurs de risque du cancer, participer à la conduite d'un véhicule autonome. Cette phase dite *d'apprentissage* ou *d'entraînement* est généralement réalisée préalablement à l'utilisation pratique du modèle. La seconde phase correspond à la mise en production : le modèle étant déterminé, de nouvelles données peuvent alors être soumises au modèle afin d'obtenir le résultat correspondant à la tâche souhaitée. Cette phase est dite de *test* et permet de nous renseigner sur la qualité du modèle à estimer un paramètre, ou à identifier la tache souhaitée. Durant la phase d'apprentissage, si les observations sont étiquetées (c'est-à-dire que la réponse à la tâche est connue pour ces données), on parle d'apprentissage supervisé. Il s'agit d'une classification ou classement si les étiquettes sont discrètes, ou d'une régression dans le cas où les étiquettes sont continues. Dans le cas le plus général, sans étiquette, on cherche à déterminer la structure sous-jacente des données et il s'agit alors d'apprentissage *non-supervisé*.

Le problème de l'apprentissage statistique supervisé se décrit comme étant une variable Y à expliquer, décrite par n individus dont on connait p variables explicatives synthétisées dans X. Connaissant un ensemble d'apprentissage $D_{Train} = (X_1, Y_1), ...(X_n, Y_n)$, on cherche

ϕ fonction des p prédicteurs telle que la variable Y s'explique au mieux en fonction des p prédicteurs :

$$Y = \phi(X)$$

Ici l'apprentissage est supervisé puisqu'il est conditionné par la donnée d'étiquettes (labels, valeurs, etc.) pour chacun des n individus : les Y_i. Voici quelques exemples classiques de méthodes d'apprentissage supervisé : modèle de régression simple ou multiple, arbre binaire de classifications, forets aléatoires, réseaux de neurones, Support Vector Machine, k-NN., etc. Quant à l'apprentissage non supervisé, il n'existe pas de variable Y à expliquer mais on a toujours n individus dont chacun est décrit par p variables. L'objectif consiste à rechercher une taxonomie (caractéristiques communes) des observations. Parmi les méthodes d'apprentissage *non-supervisé*, on cite *clustering hiérarchique*, et *k-means*.

Plus concrètement, Gregorutti (2015) définit un problème d'apprentissage supervisé comme suit :

Considérons un couple $(\mathbf{X}, Y)$ de variables aléatoires à valeurs dans $\mathbb{R}^p \times \mathcal{Y}$ dont la loi jointe $P_{(\mathbf{X},Y)}$ est inconnue. L'apprentissage supervisé consiste à estimer le lien entre $\mathbf{X} = (X_1, ..., X_p)$ (le vecteur des covariables) et Y (la variable de sortie), c'est-à-dire une fonction mesurable f définie sur $\mathbb{R}^p$ et à valeurs dans $\mathcal{Y}$.

L'erreur commise par une fonction f pour la prédiction de Y est donnée par :

$$R(f) = \mathbb{E}l(Y, f(\mathbf{X}))$$

où, l est une fonction de perte fixée. L'application R, appelée risque, mesure l'écart moyen entre Y et sa prédiction $\mathbf{f}(\mathbf{X})$. La meilleure fonction de prédiction f^* est alors celle qui minimise le risque sur la classe $\mathcal{F}$ des fonctions définies sur $\mathbb{R}^p$ et à valeurs dans $\mathcal{Y}$, c'est à dire :

$$f^* \in \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ R(f).$$

Cependant, à défaut de pouvoir calculer une telle fonction, ne connaissant pas la loi $P_{(\mathbf{X},Y)}$, nous pouvons l'estimer à partir d'un échantillon $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)\}$ de n observations indépendantes et identiquement distribuées et de loi $P_{(\mathbf{X},Y)}$, où $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$. Autrement dit, il s'agit de trouver une solution au problème de minimisation de l'estimateur empirique $\widehat{R}(f)$ de $R(f)$ en se restreignant à une sous-classe $\mathcal{C}$ de $\mathcal{F}$:

$$\begin{aligned} \widehat{f} &\in \underset{f \in \mathcal{C}}{\operatorname{argmin}}\ \widehat{R}(f) \\ &= \underset{f \in \mathcal{C}}{\operatorname{argmin}}\ \frac{1}{n} \sum_{i=1}^{n} l(Y_i, f(\mathbf{X}_i)). \end{aligned}$$

L'estimateur ainsi optimisé fournit une prédiction $\widehat{f}(\mathbf{X}_{n+1})$ pour une nouvelle observation $\mathbf{X}_{n+1}$, prédiction que l'on espère proche de la vraie valeur $\mathbf{Y}_{n+1}$. Autrement dit, il s'agit de trouver $\hat{f}$ de sorte que $\widehat{R}(f)$ est proche de $R(f^*)$. Ce principe, appelé Minimisation du Risque Empirique, a été formalisé par Vapnik (Vapnik, 1995; Vapnik et Vapnik,

1998).

Dans le cadre de cette thèse , les deux problèmes d'apprentissages que nous considérons sont respectivement la régression où $\mathcal{Y} = \mathbb{R}$ et la classification binaire lorsque $\mathcal{Y} = (0, 1)$. En régression, la fonction de perte généralement utilisée est la perte quadratique $l(Y, f(\mathbf{X})) = (\mathbf{y} - f(\mathbf{x}))^2$ et la fonction à estimer est $f^* = \mathbb{E}[\mathbf{Y}|\mathbf{X} = \mathbf{x}]$. L'erreur de prédiction ou erreur quadratique moyenne est de :

$$\mathbb{E}[(\mathbf{y} - f(\mathbf{x}))^2] \tag{3.2}$$

Si l'écriture 3.2 semble un peu technique, on peut se contenter de l'interprétation : la moyenne du carré de l'écart de la variable à expliquer est $\mathcal{Y}$ et son estimation $f(X)$ dans la population. On fera souvent appel à sa version approchée dite erreur d'entraînement calculée sur le jeu de données de test ou de validation $(x_1, y_1), ..., (x_n, y_n)$ définie par :

$$\left(\frac{1}{n}\right) \sum_{i=1}^{n} ((\mathbf{y} - f(\mathbf{x}))^2)$$

Dans ce cas-ci, on admet qu'on a déjà déterminé une fonction de régression et qu'on souhaite calculer son *erreur* d'entraînement. La valeur de la variance expliquée par le modèle de régression est complémentaire $(1 - erreur)$ de l'erreur de prédiction. La part de la variance expliquée correspond au ratio de la variance expliquée et de la variance totale et exprime la performance du modèle de régression à estimer ou à prédire une valeur de la variable à expliquer($\mathbf{Y}$) à travers les variables explicatives ($\mathbf{X}$).

Dans le cas de la classification, l'objectif est d'estimer les probabilités à posteriori :

$$\Pr[\mathbf{Y} = 1|\mathbf{X} = x] \quad ou \quad \Pr[\mathbf{Y} = 0|\mathbf{X} = x]$$

Pour transformer ces probabilités a posteriori en une fonction de décision f^*, considérée comme la fonction de la classification, on doit choisir un seuil $\mathbf{T}$ (probabilité d'acceptation) de sorte que la fonction de décision s'écrit comme suit :

$$f^*(\mathbf{x}) = \begin{cases} 1 & \text{si } \Pr[\mathbf{Y} = 1|\mathbf{X} = x] \geq \mathbf{T} \\ 0 & \text{sinon} \end{cases} \tag{3.3}$$

L'équation 3.3 nous montre qu'en fonction de seuil choisi pour $\mathbf{T}$, des objets à classer pourraient changer de classe suite au changement de la formule de la fonction de décision. Ainsi, Pour un $\mathbf{T}$ donné, nous pouvons calculer deux indicateurs de performance du classifieur, le *TPR* (True Positive Ratio), et *FPR* (False Positive Ratio). Si un objet appartenant à la base à la classe positive et qu'il a été détecté positif, il rentre dans la catégorie des *vrais positifs* (VP) tandis que s'il a été détecté négatif, il est compté dans la catégorie des *faux négatifs* (FN). De même, quand un objet de la classe négative est détecté positivement, il est compté dans la catégorie des *faux positifs* (FP) mais s'il est détecté négativement, il rentre dans la catégorie des *vrais négatifs* (VN). Ainsi, l'indicateur dit de *sensibilité* (rappel ou TPR) correspond le rapport des *vrais positifs* sur la somme des *vrais positifs* et des *faux négatifs* ($TPR = VP/(VP + FN)$). L'indicateur dit de *spécificité* est égal à $1 - FPR$ avec $FPR = FP/FP + VN$. Comme illustré dans la figure 3.9 à gauche, pour un $\mathbf{T}$ variable, nous pouvons avoir des TPR et FRP différents. Ainsi quand $\mathbf{T}$ tend vers 1, le classifieur est qualifié de *trop exigeant*, ce qui nous donne un TPR faible mais aussi un FPR faible. Par contre, si $\mathbf{T}$ tend vers 0, le classifier est jugé de *trop laxiste*, on

obtient donc un TPR élevé mais aussi un FPR élevé. A la limite pour un $\mathbf{T} = 1$,on a $TPR = PFR = 0\%$, et pour un $\mathbf{T} = 0$, on obtient $TPR = PFR = 100\%$.

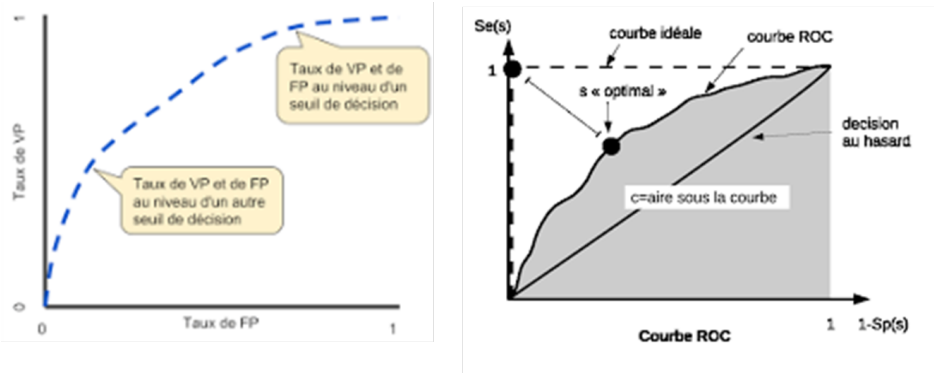


FIGURE 3.9 – Illustration de la courbe de ROC à travers le seuil sur la fonction de décision. A gauche, on observe deux valeurs de seuils donnant différents TPR et TFR. A droite est illustré la courbe de ROC avec la localisation du seuil optimal **s**. Source :**https://codingbrewery.com/2019/08/11/receiver-operating-characteristic/**

Afin de remédier au problème lié au choix du seuil de **T**, on peut tracer la courbe ROC (voir figure 3.9 à droite). Cette dernière représente les performances d'un modèle de classification pour tous les seuils de classification **T**. La courbe de ROC trace le taux de vrais positifs en fonction du taux de faux positifs. Cela permet d'évaluer globalement la performance d'une méthode sans fixer de seuil. La performance du classifieur est exprimé à travers l'indicateur appelé *AUC*. L'AUC signifie *aire sous la courbe ROC*. Cette aire peut être calculée comme l'intégrale entre 0 et 1 de la fonction que représente la courbe ROC, c'est à dire la fonction $TPR = f(FPR)$. Elle fournit une mesure agrégée des performances pour tous les seuils de classification possibles. On peut interpréter l'AUC comme une mesure de la probabilité pour que le modèle classe un exemple positif aléatoire au-dessus d'un exemple négatif aléatoire. Si par exemple, ce classifieur a une AUC de 80%, cela signifie que ce dernier respectera le bon ordonnancement (celui de classer un objet positif aléatoire au dessus d'un objet négatif) à 80% des cas. En d'autre terme, l'AUC indique à quel point le classifieur respecte le bon ordre entre les objets à classifier (ici les bâtiments en fonction de leur qualité) et exprime la qualité de classification quel que soit le seuil de **T** choisi.

L'AUC présente les avantages suivants :

- L'AUC est insensible aux valeurs absolues des probabilités fournies par l'algorithme : elle mesure plus la qualité du classement des prédictions que celle des probabilités fournies par l'algorithme. En effet, les probabilités fournies par certains algorithmes sont difficilement interprétables, et ne sont pas toujours de bonnes mesures de probabilités au sens strict du terme. L'AUC permet de s'affranchir de ce défaut en se concentrant sur le classement.
- L'AUC est indépendante des seuils de classification. Elle mesure la qualité des précisions du modèle quel que soit le seuil de classification sélectionné.

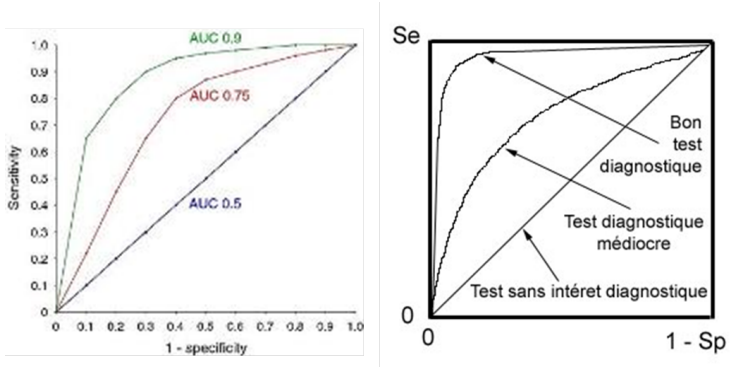


FIGURE 3.10 – Illustration de la performance du classifieur à travers l'AUC. A gauche, on observe trois courbes de ROC avec leurs AUC. Plus la courbe est tirée vers le haut et à gauche, plus l'AUC est élevé. A droite on remarque la courbe la plus haute est celle qui admet la plus haute performance de classification. Source :**http://www.adscience.fr/uploads/ckfiles/files/html_files/StatEL/statel_courbe_ROC.htm**

Toutefois, ces deux avantages comportent des limites qui peuvent réduire la pertinence de l'AUC dans certains cas d'utilisation :

- L'insensible aux valeurs absolues des probabilités n'est pas toujours souhaitable. Par exemple, nous avons parfois besoin d'obtenir des probabilités précisément calibrées, ce que l'AUC ne permet pas de déterminer.
- L'indépendance vis-à-vis des seuils de classification n'est pas toujours souhaitable lorsque des disparités importantes de coût existent entre les faux négatifs et les faux positifs, il peut être essentiel de minimiser l'un des types d'erreur de classification. Par exemple, dans un contexte de détection de spam il sera probablement préférable de minimiser en priorité les faux positifs (même si cela entraîne une augmentation significative des faux négatifs). L'AUC n'est pas un critère à retenir pour ce type d'optimisation.

3.4.2.2 Compromis biais-variance pour la performance d'un modèle d'apprentissage

En régression, afin d'obtenir une bonne performance de prédiction, le modèle d'apprentissage doit faire attention à ce qu'on qualifie du *compromis biais-variance*. Pour illustrer le compromis biais-variance, on s'intéresse aux performances d'un estimateur d'un modèle d'apprentissage. La performance de l'estimateur se mesure tout d'abord par sa capacité à fournir des prévisions assez performantes sur un jeu d'entrainement (bonne précision = biais faible). Puis elle se mesure sur sa capacité à conserver des performances proches quand le jeu de données sur lequel il est appliqué change (variance faible).

Plus précisément, quand un modèle d'apprentissage est trop simple, il ne permet pas de modéliser la réalité de manière satisfaisante. Son biais est élevé. En revanche, comme il est simple, il est assez déterministe et prédictible, un test sur des données différentes va donner un résultat très similaire en termes de performance. Sa variance est faible. A l'inverse, quand le modèle est trop complexe, il peut en général très bien s'ajuster aux données qui lui sont fournies, puisqu'il a beaucoup de degrés de liberté. Son biais est très

faible. Mais comme il a beaucoup de degrés de liberté, il risque de capturer aussi le bruit des données. Sa variance risque donc d'être élevée. On parle de sur-apprentissage. Entre ces deux situations extrêmes, il y a un optimum, avec un nombre modéré de degrés de liberté, un biais et une variance réduites. En d'autre termes, le biais correspond à l'erreur d'entrainement tandis que la variance correspond à l'erreur de test, c'est-à-dire l'erreur observée pour des données test nouvelles. L'erreur totale est l'erreur moyenne commise par une méthode d'apprentissage statistique pour prédire une réponse sur une nouvelle observation, qui n'a pas été utilisée pour ajuster le modèle. En revanche, l'erreur d'entraînement peut être facilement calculée en appliquant la méthode d'apprentissage sur les données d'entraînement. Bien que les deux erreurs soient une erreur de prédiction, l'erreur d'entraînement est souvent bien différente de l'erreur de test, et en particulier, l'erreur d'entraînement peut sous-estimer grandement l'erreur de prédiction globale du modèle, et celle-ci est mieux estimée sur des données test indépendantes du jeu d'entraînement. Ainsi pour mieux déterminer l'erreur de prédiction, nous nous basons sur le calcul de l'erreur de test.

Par conséquent, la recherche du compromis biais-variance optimal garantit la meilleure voie pour l'estimation de l'erreur de prédiction d'un modèle de régression. La figure 3.11 illustre le calcul deux courbes correspondant à l'erreur de prédiction calculée de deux manières différentes en fonction de la complexité du modèle. Les modèles complexes ont souvent une grande variance tandis que les modèles simples ont un grand biais. La courbe en vert illustre le phénomène de sur-apprentissage : l'optimisme de l'erreur de prédiction calculée sur le jeu de données d'entraînement. La courbe en rouge illustre la nécessité d'utiliser un jeu de données test qui n'a pas servi à l'estimation (entrainement) du modèle. La complexité optimale du modèle correspond au compromis biais-variance.

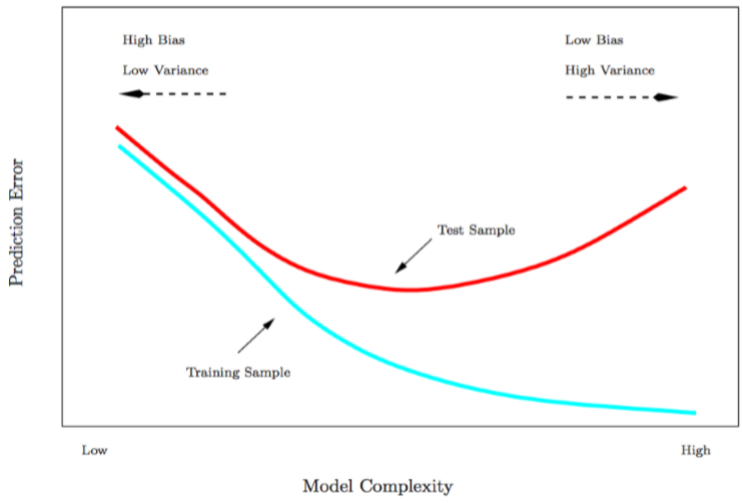


FIGURE 3.11 – Illustration du compromis biais-variance à travers le calcul de l'erreur de prédiction en fonction de la complexité du modèle. Source : **https://www.aspexit.com/comment-valider-un-modele-de-prediction/**

Afin de trouver un nombre modéré de paramètres, et de surcroit atteindre le compromis biais-variance, nous faisons appel aux méthodes de régularisation (méthode LASSO)

ou/et à des méthodes destinées à conserver un nombre restreint des variables explicatives dans un modèle de régression (méthode AIC ou BIC).

3.4.2.3 Sélection de variables

La sélection des variables qui expliquent la variable réponse y est une question cruciale en régression linéaire multiple et en classification. La première raison est numérique, il se trouve que la matrice $X^T X$ qui intervient dans la solution des moindres carrés devient non-inversible dès que le nombre de variables explicatives p est supérieur au nombre d'individus n. Il est donc indispensable de désigner un sous-ensemble de variables explicatives liées à y et à écarter les variables explicatives qui n'apportent aucune information sur y. Certaines méthodes permettent de choisir le sous-ensemble de variables explicatives à conserver dans le modèle de régression. Ces méthodes reposent sur une recherche dite de *pas à pas* (*stepwise* en anglais) pour trouver le modèle qui maximise le critère d'information qui réalise un compromis entre l'ajustement aux données et la complexité du modèle. Les critères d'information que nous utilisons dans le cadre de cette thèse sont AIC (*Akaike Information Criterion*) et BIC (*Bayesian Information Criterion*). Ainsi, dans une logique de recherche *backward*, en partant du modèle complet incluant toutes les variables comme modèle d'initialisation, nous retirons à chaque itération la variable dont l'exclusion du modèle réalise le meilleur gain en termes de critère d'information. Nous nous arrêtons lorsque aucune exclusion n'améliore le critère ou lorsque le modèle vide est atteint.

Pour atteindre un équilibre entre l'ajustement et la parcimonie et déterminer le modèle de régression optimal pour la perte d'information, les critères AIC et BIC sont adaptés à nos besoins. Dans l'étude réalisée dans Yang (2005), la comparaison faite sur l'utilisation de l'AIC et du BIC, suggère d'employer la méthode AIC. En effet, lorsqu'on souhaite faire des prévisions, le modèle issu de l'AIC doit être utilisé mais si le but de l'apprentissage est d'expliquer, le meilleur modèle est obtenu en utilisant le critère BIC. Le critère AIC donne un modèle efficace tandis que le modèle issu du BIC récupère le *vrai* modèle (sous réserve que ce dernier fasse partie des modèles à comparer). Ainsi, dans notre cas, où l'objectif d'apprentissage est d'estimer ou de prédire une valeur de la variable dépendante, nous optons au final pour l'utilisation du critère AIC pour la sélection des variables les plus significatives dans le modèle de régression multiple.

Toutefois, la sélection basée sur la méthode AIC peut faire subvenir certains problèmes. En effet, un grand nombre de variables explicatives avec une forte corrélation est susceptible de rendre impossible la mise en place des procédures pas à pas pour la sélection de variables. Un autre problème est dû à la discontinuité de la sélection de variables par méthode de *stepwise*, c'est-à-dire une modification infinitésimale des données impacte considérablement le sous-ensemble de variables sélectionnées. On parle aussi d'*instabilité* de la sélection *stepwise* des variables en grande dimension.

Face au problème de grande dimension du modèle de régression, les méthodes de régularisation apportent une meilleure réponse à la problématique de sélection de variables. Cette réponse repose sur les remarques suivantes :

- Si les coefficients β_j ne sont pas contraints, ils peuvent prendre de très grandes valeurs et donc entrainer une grande variance.

- Pour contrôler la variance, il faut contrôler la taille des coefficients de β. Cette approche pourrait réduire sensiblement l'erreur de prédiction.

On cherche à minimiser la solution du problème des moindres carrés suivante :

$$\underset{\beta \in \mathbb{R}^p}{min} \; ||Y - \beta X||_2$$

Algorithme et Interprétation de la méthode LASSO

Si on considère une régularisation de type LASSO on obtient un modèle qui s'écrit comme suit :

$$\min_{\beta \in \mathbb{R}^p} \; ||Y - \beta X||_2 \; + \; \lambda ||\beta||_1 \tag{3.4}$$

où le 1er terme ($\min_{\beta \in \mathbb{R}^p} \; ||Y - \beta X||_2$) est appelé *attache aux données* et le second terme ($\lambda ||\beta||_1$) est *régularisation ou pénalité Lasso*

On peut montrer que 3.4 est la forme lagrangienne d'un problème d'optimisation sous contrainte :

$$\underset{\beta \in \mathbb{R}^P}{min} \; ||Y - \beta X||_2 \; sous \; la \; contrainte \; ||\beta||_1 \leqslant t$$

avec une équivalence bijective (dont la formulation exacte dépend des données) entre λ et t (t est d'autant plus petite que λ est grande).

Pour une valeur de t donnée, la zone autorisée est une ligne de niveau de la norme L_1 sous la forme d'un losange. La forme de la zone découle du fait que la L_1 est la seule norme qui soit à la fois non-dérivable et convexe. La non-dérivabilité autorise l'existence des sommets *anguleux* sur la frontière de la forme. Cela permet l'annulation des coefficients des variables secondaires (variables explicatives moins significatives) comme illustré sur la figure figure 3.12 .

Par ailleurs avec toutes les normes L_p, pour $p = 1$, le problème est convexe, ce qui facilite la résolution numérique, mais pour $p > 1$, il n'y a pas d'*incitation* à annuler un maximum de coefficients assurant ainsi la praticabilité d'une résolution numérique du problème.

Sur la figure 3.12, est illustrée à gauche et en bleu, la *pénalisation lasso* sous forme de losange en norme L_1, et a droite en bleu, est représentée la*pénalisation ridge* sous forme d'un cercle avec une pénalité en norme L_2. En rouge les niveaux de β des moindres carrées. On observe que *l'estimateur lasso* raccroche mieux que *l'estimateur ridge* grâce à sa forme anguleuse, annulant ainsi les valeurs des β (ici β_1) tandis que *l'estimateur ridge* minimise les deux coefficients, sans toutefois les annuler (Friedman et al., 2001).

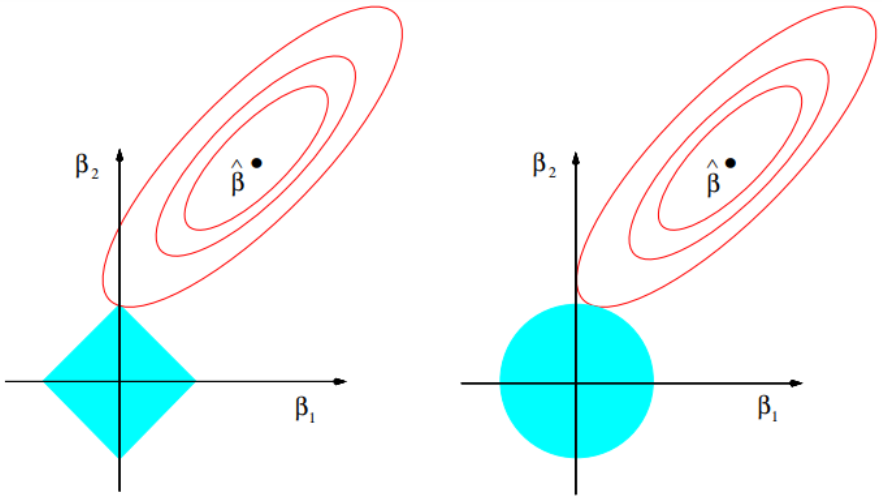


FIGURE 3.12 – Illustration de la solution des moindres carrées appliqué par une pénalisation lasso et une pénalisation ridge. Source : (Friedman et al., 2001)

Ainsi avec la méthode LASSO, certains coefficients sont mis à zéro, des telles solutions, avec plusieurs coefficients qui sont identiquement nuls, sont dites *sparses*. La pénalité effectue donc une sorte de sélection continue des variables. L'estimateur β_{lasso} avec un λ donné, obtenu en minimisant l'équation 3.4 est appelé *lasso pour Least Absolute Shrinkage and Selection Operator*.

Ainsi, pour pénaliser davantage le modèle de régression et ne retenir que les variables les plus importantes, nous peuvons réaliser une régression avec la méthode LASSO. Les variables ayant un coefficient de régression égal à zéro après le processus de régularisation sont *de facto* exclues du modèle, tandis que celles dont les coefficients de régression sont non nuls demeurent significativement associées à la variable dépendante. Cette démarche fournit un modèle fondé sur un ensemble strict de variables explicatives, et donc facilement interprétable. Par ailleurs, l'exclusion de certaines variables peu informatives peut contribuer à augmenter les capacités de généralisation du modèle ainsi obtenu en lui permettant de se prémunir des effets néfastes du sur-apprentissage.

La méthode LASSO a été utilisée à des nombreuses reprises dans le domaine de l'information géographique. C'est le cas de l'étude Inoue et al. (2018) sur la segmentation géographique du marché immobilier. En utilisant une méthode LASSO de fusion généralisée, les auteurs cherchent à extraire les variables les plus importantes parmi les paramètres régionaux d'un modèle de prix.

Comme mentionné précédemment, la méthode LASSO est une minimisation de l'erreur de régression avec la fonction des moindres carrées ordinaires notée *MCO* et avec une pénalité de norme L1. Cela impose de rechercher dans un espace en forme de losange, une solution réalisant une valeur minimale de *MCO*.

Dans le cas de la méthode LASSO, si l'on reconsidère le problème de minimisation donnée par l'équation 3.4, nous pouvons observer que la fonction à minimiser f est constitué de la somme d'élément en norme L_2 et d'un élément de pénalisation en norme

L_1, conférant à la fonction une propriété de convergence mais non-différentiable. Pour s'assurer que le problème converge vers une solution, nous devons transformer la fonction en une somme des fonctions séparables qui sont elles différentiables en toute valeur de β_j. Par la suite, nous utilisons un processus itératif de sorte à calculer des valeurs approchées de β_j afin d'obtenir un minimum pour la fonction à minimiser f. Dans la littérature, nous relèvons deux algorithmes pour résoudre le problème d'optimisation : *cyclical cordinates descent* et l'algorithme *proximal gradient descent.*

Pour converger vers la solution, la méthode *cyclical cordinates descent* fait appel à un algorithme itératif qui peut se comprendre comme étant un algorithme minimisant successivement la fonction f le long des axes de coordonnées. Pour chaque itération, l'algorithme de *cyclical cordinates descent* minimise f le long d'une seule direction, c'est-à-dire en fixant tous les β_j sauf un. L'algorithme fait varier la direction courante dans une boucle de taille égale au nombre des β. L'objectif consiste à trouver une solution (approchée) de $\underset{\beta}{\text{argmin}} f(\beta)$. L'algorithme fonctionne donc comme suit :

Entrées : f, nombre d'itérations K

Initialisation $k = 0$ *et* $\beta^{(0)} = 0 \ \in \ \mathbb{R}^p$

pour $k = 0, 1, 2, ...K-1$ **faire**

$\beta_1^{(k+1)} \leftarrow \underset{\beta_1 \in \mathbb{R}}{\text{argmin}}\ f(\beta_1, \beta_2^{(k)}, \beta_3^{(k)}, ..., \beta_{p-1}^{(k)}, \beta_p^{(k)})$

$\beta_2^{(k+1)} \leftarrow \underset{\beta_2 \in \mathbb{R}}{\text{argmin}}\ f(\beta_1^{(k+1)}, \beta_2, \beta_3^{(k)}, ..., \beta_{p-1}^{(k)}, \beta_p^{(k)})$

$\beta_3^{(k+1)} \leftarrow \underset{\beta_3 \in \mathbb{R}}{\text{argmin}}\ f(\beta_1^{(k+1)}, \beta_2^{(k+1)}, \beta_3, ..., \beta_{p-1}^{(k)}, \beta_p^{(k)})$

.

.

.

$\beta_p^{(k+1)} \leftarrow \underset{\beta_p \in \mathbb{R}}{\text{argmin}}\ f(\beta_1^{(k+1)}, \beta_2^{(k+1)}, \beta_3^{(k+1)}, ..., \beta_{p-1}^{(k+1)}, \beta_p)$

Fin de boucle sur k

Sorties : $\beta^{(K)}$

Critères d'arrêts sont : *itérés stables ou objectifs stables,etc.*

On doit visiter toutes les coordonnées régulièrement pour assurer la convergence d'où le choix d'un parcours cyclique de l'algorithme.

Sur la figure 3.13, nous observons que la convergence vers la solution à l'aide de descente de coordonnées cyclique se fait d'en haut vers le bas suivant les deux directions *W0* et *W1*. A chaque étape, nous recherchons le minimum sur une direction, puis sur l'autre tout en maintenant fixée l'autre coordonnée.

Pour pouvoir appliquer l'algorithme de *cyclical cordinates descent*, une légère modification s'impose. La fonction à minimiser doit être différentiable. On doit transformer le problème en celui d'un problème d'optimisation dit *proximal* via la méthode dite *proximal gradient descente* rendant ainsi la fonction à la fois convexe et différentiable et s'écrivant

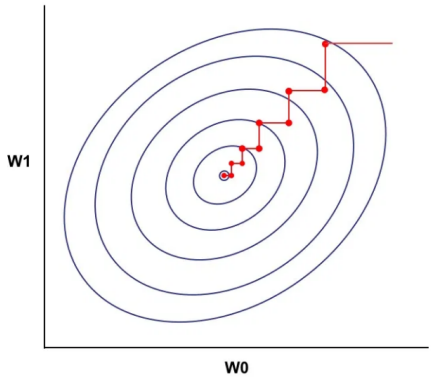


FIGURE 3.13 – Illustration du principe de la convergence avec la méthode de descente de gradient. Source : **http://www.adeveloperdiary.com/data-science/machine-learning/introduction-to-coordinate-descent-using-least-squares-regression/**

de la forme suivante :

$$f(\beta) = \frac{1}{2}||Y - X\beta||^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

A chaque étape, on minimise β_j avec les autres $\beta_k (k \neq j)$ fixes :
$\hat{\beta} = \arg\min_{\beta_j \in \mathbb{R}} f(\beta_1, ...\beta_p)$

$= \arg\min_{\beta_j \in \mathbb{R}} \frac{1}{2}||y - \sum_{k \neq j} \beta_k x_k - x_j \beta_j||^2 + \lambda \sum_{k \neq j} |\beta_k| + \lambda |\beta_j|$

Ce qui nous donne après développement :

$$\tilde{\beta}_j \leftarrow S\Big(\frac{1}{n} \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)}), \lambda\Big)$$

où $\tilde{y}_i^{(j)} = \tilde{\beta}_0 + \sum_{k \neq j} x_{ik} \tilde{\beta}_k$ est la valeur estimée de y_i sans la contribution de x_{ij} , et donc $y_i - \tilde{y}_i^{(j)}$ est le résidu partiel pour l'ajustement de β_j. En raison de la normalisation, le terme $\frac{1}{n} \sum_{i=1}^{n} x_{ij}(y_i - \tilde{y}_i^{(j)})$ est le coefficient des moindres carrés simples lors de l'ajustement de ce résidu partiel à x_{ij}. Ce coefficient n'est rien d'autre que le résidu moyen relatif à l'estimation de β_j et sans la contribution des x_{ij}.

La fonction S est définie comme suit : :

$$S(z, \gamma) = \left\{ \begin{array}{ll} z - \gamma & \text{si } z > 0 \text{ et } \gamma < |z| \\ z + \gamma & \text{si } z < 0 \text{ et } \gamma < |z| \\ 0 & \text{sinon} \end{array} \right\}$$

Si l'on revient à la logique de l'algorithme, au fil des itérations, on cherches un λ optimal donnant les solutions les plus *stables* qui minimisent l'erreur de prédiction. Pour cela on se fixe une plage des valeurs de λ qui correspondent aux différentes itérations de l'algorithme.

Dans la pratique, nous utilisons une procédure de validation croisée qui subdivise le jeu de données en deux parties (pour l'entrainement et pour la validation). L'erreur de prédiction est ensuite calculée à à chaque itération sur λ (il y a aussi des itérations pour la descente de gradient) sur les données de validation. Au cours des itérations, l'algorithme converge au moment où on obtient un λ_{min} qui rend stable le RMSE sur l'erreur de prédiction. Par souci d'interprétabilité, on peut s'arrêter sur la valeur λ_{1se} qui correspond à un écart-type au-dessus de la valeur de λ_{min}. Avec λ_{1se}, on pénalise un peu plus le modèle de régression au profit de l'annulation certains coefficients de β, ceci étant en accord avec la stratégie de la méthode LASSO qui cherche à réduire le nombre des variables explicatives.

3.4.2.4 Estimation de l'erreur de prédiction

Si les méthodes de régularisation permettent de réduire l'erreur de prédiction, d'autres tentent de l'estimer. Ce sont les méthodes de validation croisée et de bootstrap. Ces méthodes ré-ajustent le modèle sur des échantillons issus de l'échantillon d'apprentissage dans le but d'obtenir des informations supplémentaires sur ce modèle. Par exemple, ces méthodes fournissent des estimations de l'erreur sur des ensembles de test, le biais et la variance des estimations de paramètres. Plus précisément, la validation croisée permet d'estimer l'erreur de test (l'erreur d'entraînement on l'obtient toujours facilement à l'issue de la procédure d'entraînement) tandis que le bootstrap sert à estimer l'incertitude flottant autour d'une grandeur aléatoire quelconque, dont en particulier l'erreur de test, quand on a un nombre réduit de données.

La validation croisée est couramment utilisée pour estimer l'erreur de test. L'estimation de l'erreur peut être utilisée pour choisir le meilleur modèle (la meilleure méthode d'apprentissage), ou approcher l'erreur de prédiction du modèle finalement choisi. L'idée est de diviser les données en K groupes de même taille comme illustré dans la figure 3.14. Nous laissons le k−ème bloc de coté, puis nous entrainons le modèle sur le reste des blocs, et enfin nous le testons sur le bloc laissé de côté. Nous répètons l'opération en laissant de côté le bloc $k = 1$, puis $k = 2$,...jusqu'au bloc $k = K$. Puis nous moyennons les performances de validation(k erreur de test) sur les k étapes.

Ainsi, pour chacune des observations, nous obtenons une prédiction $\hat{y}_i = \hat{m}(xi)$ ou $\hat{g}(xi)$au moment où i est dans le groupe mis de côté. Nous comparons alors ces prédictions aux étapes comme pour l'erreur de test :

$$\text{err(K)}= \left(\frac{1}{n}\right) \sum_{i=1}^{n}(y_i - \hat{y}_i) \qquad \text{ou} \qquad \text{err(K)}= \left(\frac{1}{n}\right) \sum_{i=1}^{n} 1\{y_i - \hat{g}(xi)\}$$

Nous avons donc, l'erreur de test(ou de prédiction) qui correspond à la moyenne des erreurs calculées à chaque étape de k ($err(k)$).

Dans le cas extrême, avec $K = n$ données, nous partageons en n groupes les observations, il s'agit du *leave-one-out cross validation (LOOCV)*. On dit que LOOCV ne secoue pas assez les données. En effet, les règles de classification $\hat{g}$ ou les fonctions de régression inférées $\hat{m}$ avec $(n - 1)$ données sont très corrélés les uns aux autres. L'erreur d'échantillonnage n'est plus visible, autrement dit la variabilité de l'estimation de la fonction. C'était pourtant tout l'intérêt de la validation croisée. En général, on choisit $K = 5$

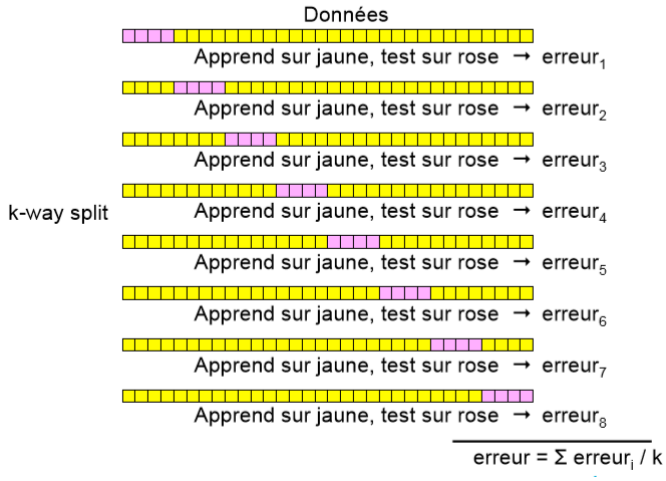


FIGURE 3.14 – Processus de la validation croisée. Source : **https://docplayer.fr/2846756-Introduction-a-la-fouille-de-donnees-khai-thac-du-lieu-cours-n-2-cour.html**

ou $K = 10$ blocs pour le processus de la validation croisée.

Par ailleurs, la méthode de bootstrap se base sur des simulations stochastiques, comme les méthodes de Monte-Carlo, les méthodes numériques bayésiennes, à la différence près que le bootstrap ne nécessite pas d'autre information que celle disponible sur les individus de l'échantillon original. Plus précisément, et c'est le sens du terme *ré−échantillonnage*, un bootstrap consiste à créer des *nouveaux échantillons* statistiques par tirage avec remise, à partir de l'échantillon de base. On remarquera donc que les échantillons bootstrap se ressemblent en grande partie, fluctuant autour de l'échantillon initial. La méthode de bootstrap est utilisée dans ce cadre pour estimer l'incertitude de la grandeur en question, l'erreur de test, sous l'hypothèse qu'à l'infini l'estimation de cette grandeur sur des échantillons bootstrap, tend asymptotiquement de sa valeur de base calculée sur l'échantillon de base. Ce qui est plus intéressant en revanche avec la méthode de bootstrap, c'est la possibilité de calculer l'incertitude associé à chaque valeur de l'erreur de test calculée, issue de la série d'échantillons bootstrap.

Pour le cas d'étude d'une régression, à une étape donnée (k−ème étape) dans le processus validation croisée, un échantillon de base est composé d'un partie de données d'entrainement et une autre partie de données de test. Un échantillon bootstrap est formé d'une sélection de certains éléments sur les deux parties de l'échantillon de base distinctement. Ainsi, l'erreur de test calculée conjointement sur l'échantillon d'entrainement et l'échantillon de test dans le processus de validation croisée, peut être recalculer à plusieurs reprises (sur 100 échantillons bootstrap) pour chaque itération de la validation croisée, afin de déterminer une valeur d'incertitude (écart−type) de la valeur de l'erreur de test. Quand la validation croisée cherche à calculer une valeur moyenne sur l'erreur de test, le bootstrap estime une valeur moyenne sur l'incertitude sur l'erreur de test afin d'apporter plus de précision sur la détermination de l'erreur de test, et ce à chaque itération de la validation croisée. La combinaison de deux méthodes permet d'obtenir une estimation de l'erreur de test ou de prédiction du modèle accompagnée d'une valeur d'incertitude (écart-type de l'erreur de test). A terme, avec le bootstrap, nous pouvons déterminer l'intervalle

de confiance dans lequel varie la valeur de l'erreur de prédiction.

3.4.3 Étude d'une auto-corrélation spatiale

3.4.3.1 Définition et principe de l'auto-corrélation spatiale

L'auto-corrélation spatiale est *la corrélation d'une variable avec elle-même (auto-corrélation) attribuable à l'organisation géographique des données (spatiale)* (Griffith, 1989). Il s'agit d'un phénomène statistique qui fait dépendre les valeurs d'une variable, des valeurs de cette même variable en des points voisins dans l'espace. Autrement dit, les valeurs des individus voisins dans l'espace sont dépendantes les unes des autres. Lorsque les valeurs des individus voisins tendent à se ressembler, on parle d'auto-corrélation spatiale positive. Lorsqu'au contraire les valeurs des individus voisins sont dissemblables, on parlera d'auto-corrélation spatiale négative. En l'absence d'auto-corrélation spatiale, on peut considérer que la répartition spatiale des observations est aléatoire. Les mesures d'auto-corrélation spatiale sont donc directement dépendantes de la définition du voisinage retenu (contiguïté, distance, plus proche voisins).

On appelle voisinage l'ensemble de ce qui est proche, rapproché, contigu défini selon une définition basée soit sur une distance, soit une relation de contiguïté. La distance mesure une séparation, un écart entre deux lieux, exprimée dans une métrique tandis que la contiguïté mesure une relation topologique. Ainsi, on peut recourir à des relations topologiques pour exprimer les contiguïtés entre objets ou lieux qui sont voisins, soit parce qu'ils ont une frontière commune s'il s'agit de mailles ou de zones, soit parce qu'ils sont reliés par une ligne s'ils correspondent à des nœuds dans un réseau. Dans chacun des cas, la contiguïté est définie de manière binaire (pour une paire de lieux donnés, elle est notée 0 si les objets ou les lieux ne sont pas contigus, 1 s'ils le sont). L'information sur les contiguïtés peut être consignée dans une matrice de contiguïté où résumée par un graphe de contiguïté. Choisir des relations de contiguïté pertinentes peut être complexe, en particulier quand les contiguïtés sont mesurées par rapport à un maillage. Il faut dans ce cas définir ce qui est appelé frontière commune (côté, angle ou encore, nœud, face). Par ailleurs, pour tenir compte du fait que des mailles séparées apparaissent cependant suffisamment proches pour conserver des possibilités de contact ou d'interaction, on peut définir des contiguïtés à différents ordres (de l'ordre 1 à l'ordre n ; à l'ordre 2 par exemple, la contiguïté entre mailles existe si, et seulement si, on traverse exactement deux frontières pour aller de l'une à l'autre par le plus court chemin). En généralisant le voisinage peut être évalué par une combinaison de distances et de relations topologiques.

La fonction de voisinage permet de construire la structure spatiale respectant l'auto-corrélation spatiale. Toutefois, on précise que la structure spatiale et l'auto-corrélation spatiale ne peuvent pas exister indépendamment l'une de l'autre (Tiefelsdorf, 1998) car d'une part, la structure spatiale renferme l'ensemble des liens grâce auxquels le phénomène auto-corrélé va se diffuser, et d'autre part, sans la présence d'un processus auto-corrélé significatif, la structure spatiale ne peut être empiriquement observée. La distribution spatiale observée est alors considérée comme la manifestation du processus spatial sous-jacent.

L'auto-corrélation spatiale peut être le résultat d'un processus inobservé ou difficilement quantifiable qui associe des localisations différentes et qui, de ce fait, se traduit par une structuration spatiale vis-à-vis d'une variable d'étude. Des phénomènes d'interaction

(entre les décisions des agents par exemple) ou de diffusion (comme les phénomènes de diffusion technologique) dans l'espace sont autant de phénomènes qui peuvent produire de l'auto-corrélation spatiale. Ainsi, l'étude d'une auto-corrélation spatiale peut aider à identifier une dépendance spatiale entre les observations d'une variable donnée. Si une telle dépendance se confirme, il peut être pertinent, pour certaines applications, d'agréger les observations dans un voisinage dans lequel ces observations sont similaires. Dans ce cas-là, au lieu de considérer les observations d'une variable à leur à l'échelle des objets individuels (échelle micro), il est intéressant d'agréger certains objets géographiques dans un voisinage et de considérer seulement une valeur de tendance centrale (par exemple, la moyenne) sur leurs observations.

En outre, d'un point de vue statistique, de nombreuses analyses (analyse des corrélations, régressions linéaires, etc.) reposent sur l'hypothèse d'indépendance des variables. Lorsqu'une variable est spatialement auto-corrélée, l'hypothèse d'indépendance n'est plus respectée, remettant ainsi en cause la validité des hypothèses sur la base desquelles ces analyses sont menées. C'est pourquoi, on estime qu'en agrégeant des observations dans la structure spatiale sur une valeur centrale ou globale, la qualité de prédiction d'un modèle de régression est meilleure. Afin de vérifier l'existence d'une auto-corrélation, on procède au calcul d'indices d'auto-corrélation spatiale sur une structure spatiale définie au préalable.

3.4.3.2 Mesure de l'auto-corrélation spatiale

Les indices d'auto-corrélation spatiale permettent d'évaluer la dépendance spatiale entre les valeurs d'une même variable en différents endroits de l'espace et de tester la significativité de la structure spatiale identifiée. Pour la mettre en évidence, les indices prennent en compte deux critères :

- la proximité spatiale ;
- la ressemblance ou la dissemblance des valeurs de la variable pour les unités spatiales considérées.

Si les données sont agrégées suivant un découpage qui ne respecte pas le phénomène sous-jacent, on surestimera ou sous-estimera la force du lien spatial.

De manière très générale, les indices d'auto-corrélation spatiale permettent de caractériser la corrélation entre les mesures géographiquement voisines d'un phénomène mesuré. Deux indices sont principalement utilisés pour tester la présence d'auto-corrélation spatiale : l'indice de Moran et l'indice de Geary. Le premier considère les variances et covariances en prenant en compte la différence entre chaque observation et la moyenne de toutes les observations. L'indice de Geary, lui, prend en compte la différence entre les observations voisines. Dans la littérature, l'indice de Moran est souvent préféré à celui de Geary en raison d'une stabilité générale plus grande (Upton et al., 1985).

L'indice de Moran est donné par :

$$I_{Moran} = \frac{n}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

où i,j = unité spatiale ; n = nombre d'unités spatiales ; xi est la valeur de la variable dans l'unité i ; $\bar{x}$ est la moyenne de x ; et w_{ij} sont des éléments de la matrice d'interactions

spatiales, définie par des relations de contiguïté, des distances ou des frontières communes.

- H_0 : Les voisins ne co-varient pas d'une façon particulière ;
- Il y a auto-corrélation spatiale positive si $I_{Moran} > 0$

L'indice de Geary est donné par :

$$C_{Geary} = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i - x_j)^2}{2\big(\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\big)\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

où i,j = unité spatiale ; n = nombre d'unités spatiales ; xi, xj sont les valeurs de la variable dans l'unité i respectivement j ; $\bar{x}$ est la moyenne de x ; et w_{ij} sont les éléments de la matrice d'interactions spatiales, définie sous la forme de la contiguïté, les distances ou les frontières communes.

- H_0 : les différences entre voisins n'ont pas de structure particulière ;
- Il y a auto-corrélation spatiale positive si $C_{Geary} < 1$.

Le calcul des indices d'auto-corrélation spatiale a pour objectif de répondre à deux questions :

- Les valeurs prises par les observations voisines auraient-elles pu être aussi comparables (ou aussi dissemblables) par le simple fait du hasard ?
- Si tel n'est pas le cas, il y a de l'auto-corrélation spatiale : quels en sont le signe et la force ?

Répondre à la première question revient à tester l'hypothèse d'absence d'auto-corrélation spatiale pour une variable brute y.

- H_0 : absence d'auto-corrélation spatiale ;
- H_1 : présence d'auto-corrélation spatiale.

Pour mener à bien ce test, il faut préciser quelle est la distribution de la variable d'intérêt y, en l'absence d'auto-corrélation spatiale (sous H_0). Dans ce contexte, l'inférence statistique est généralement menée en considérant l'une ou l'autre des deux hypothèses suivantes :

- Hypothèse de normalité : chacune des valeurs de la variable, soit yi, est le résultat d'un tirage indépendant dans la distribution normale propre à chaque zone géographique i sur laquelle est mesurée cette variable ;
- Hypothèse de randomisation : l'inférence sur l'indice de Moran est généralement menée sous l'hypothèse de randomisation. En effet, dans le cas de l'auto-corrélation spatiale, on se fixe l'hypothèse nulle comme étant qu'il n'y a pas d'association spatiale. Puis, on réalise des permutations affectant les valeurs de la variable à des unités spatiales différentes que les leurs. Ensuite, on compare l'estimation de la statistique (l'indice de Moran) obtenue à partir des données initiales à la distribution des valeurs de l'indice de Moran obtenues en réordonnant au hasard (permutations) les données. Cela revient à calculer une sorte d'écart entre l'indice de Moran initial et l'espérance des indices de Moran issus des permutations, si bien que plus l'écart est grand, plus l'hypothèse d'auto-corrélation est soutenue. L'idée est simplement que si l'hypothèse nulle est vraie, alors toutes les combinaisons possibles de données sont équiprobables, et l'écart entre l'indice de Moran et sa moyenne sur les différentes

permutations (généralement appelé Z-score) est proche de 0. Si l'hypothèse nulle est rejetée, c'est-à-dire s'il y a de l'auto-corrélation spatiale, la valeur du Z-score doit être supérieure à 1 avec une $p - value$ (la significativité du degré d'auto-corrélation spatiale) inférieure à $0,05$.

L'espérance mathématique des indices de Moran (hypothèse de non auto-corrélation spatiale) est donnée par :

$$\mathbb{E}[I_{Moran}] = -\frac{1}{n-1}$$

Deuxième partie

Démarche et mise en œuvre

Chapitre 4

Démarche pour l'évaluation de la qualité avec des indicateurs intrinsèques

Sommaire

Dans le but d'évaluer la qualité extrinsèque d'un jeu de données OpenStreetMap en l'absence de données de référence, ce chapitre décrit une méthodologie globale sous la forme d'un workflow. Ce dernier recense l'ensemble des étapes à réaliser progressivement pour arriver à une estimation de la qualité extrinsèque d'un jeu de données quelconque avec des indicateurs intrinsèques, pour déterminer quels objets sont a priori de qualité suffisante pour être intégrés dans un référentiel ou utilisés dans des applications. Nous distinguons deux grandes phases. La 1ère phase correspond à l'ensemble des tâches que l'on réalise durant cette thèse dont l'objectif consiste à fournir une méthode pour estimer de la précision spatiale d'un jeu de données OSM. La seconde phase regroupe les étapes qu'un utilisateur quelconque voulant vérifier la qualité sur son nouveau jeu de données OSM (par exemple nouvelle zone géographique sur laquelle on souhaiterait évaluer la qualité des données), pourrait exécuter. Après l'estimation de la qualité extrinsèque, l'utilisateur se fixant un seuil de qualité, pourrait accepter les bâtiments dont la qualité estimée est suffisante, et rejeter les autres pour utiliser ces données pour une application donnée. Ainsi le workflow proposé sert d'abord à produire un modèle d'évaluation de la qualité d'un jeu de données de façon globale. Puis, une fois le modèle établi, la démarche fournit méthode d'estimation de la qualité d'un nouveau jeu de données à travers un seuil de qualité fixé selon l'application souhaité. En fin du chapitre, nous présentons la zone d'étude et les données utilisées dans nos travaux de thèse.

4.1 Méthodologie pour l'évaluation de la qualité

4.1.1 Approche globale

Nous pouvons évaluer la qualité des données de la base OSM en les comparant avec les données d'une base de référence. Cependant, en l'absence d'une telle base de référence, la précision spatiale de ces données n'est pas connue. L'objectif de nos travaux est de mettre en place un cadre d'étude permettant de déterminer la précision spatiale d'un jeu de données OSM sans devoir faire appel à une base de référence. Pour cela, nous cherchons à établir un lien statistique entre des mesures extrinsèques de qualité (calculées en confrontant les données OSM avec des données de référence), et des indicateurs intrinsèques de qualité (calculés en utilisant uniquement les objets à évaluer), pour disposer d'un modèle d'estimation des mesures extrinsèques de qualité d'un jeu de données OSM pour lequel il n'y aurait pas de référence. Le schéma suivant illustre les différentes phases à réaliser pour y parvenir.

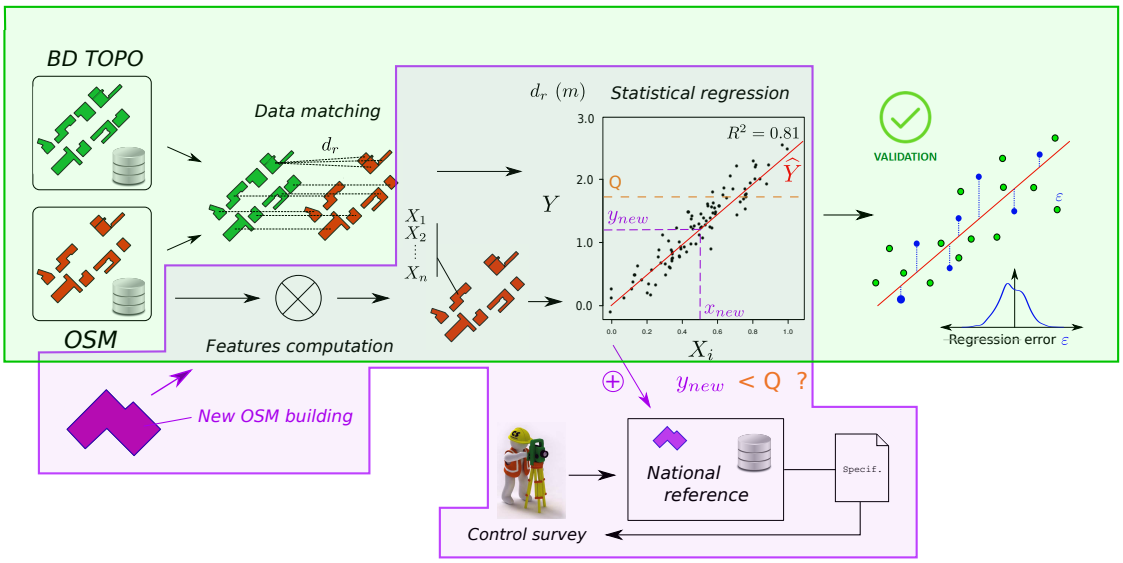


FIGURE 4.1 – méthodologie générale pour mesurer la précision spatiale d'un jeu de données sans données de référence

Dans un premier temps, nous procédons à l'extraction de deux jeux de données, l'un extrait d'un jeu de données de référence (dans cette thèse nous avons utilisé la BDTOPO de l'IGN) et l'autre jeu de données à évaluer (dans cette thèse nous avons utilisé les données de la plateforme OSM). Nous définirons plus tard le type de données extrait sur la zone d'étude. Puis s'ensuit l'étape d'appariement précédée du choix des indicateurs extrinsèques. Ce choix consiste à trouver dans la littérature scientifique un ensemble d'indicateurs qui décrivent la précision géométrique et positionnelle. Sur chaque objet des deux jeu de données, nous calculons un certain nombre de mesures. Pour chaque mesure définie et calculée, on peut comparer deux objets (un objet issu de la BDTOPO et un objet issu d'OSM) en calculant l'écart entre les valeurs que prend cette mesure pour ces deux objets. Ces écarts constituent des mesures extrinsèques de la qualité de la base OSM en prenant la base BDTOPO comme référence.

A l'étape suivante dite d'appariement de données, nous définissons un algorithme de correspondance multicritère capable d'identifier des objets homologues entre la base OSM et les données de référence. Pour chaque objet à apparier nous cherchons des candidats, puis pour tous les couples (objet-à-apparier,candidat), nous calcules les mesures extrinsèques. L'algorithme d'appariement, pour chaque objet OSM, choisit parmi les candidats (de la base BDTOPO), l'objet qui correspond le mieux sur la base des écarts calculés pour chacun des critères d'appariement. A l'issue de l'appariement, chaque objet OSM apparié porte en guise d'étiquette, les valeurs des écarts avec l'objet apparié.

Ensuite, nous procédons au choix et au calcul des indicateurs intrinsèques sur des objets de la base OSM. Cette étape consiste à définir des indicateurs intrinsèques basés sur des propriétés qui reflètent la qualité spatiale (géométrique et positionnelle) d'un objet. Ce sont les indicateurs qui sont censés être en relation avec les mesures extrinsèques (exprimant la précision spatiale).

L'étape d'après est celle de l'inférence de la qualité extrinsèque à partir de la qualité intrinsèque. Elle consiste à appliquer des techniques d'apprentissage statistique pour identifier des liens statistiques entre les indicateurs intrinsèques et des mesures extrinsèques afin d'établir un modèle permettant d'estimer une mesure extrinsèque. Cette mesure extrinsèque correspondant à un écart de position ou de géométrie entre un objet et ce qui serait son équivalent si on disposait d'une base de référence. Ces techniques permettent d'estimer une valeur de précision par régression, ou d'identifier les données dont la qualité est supérieure à un certain (bonne qualité) seuil par classification. Dans cette étape, on fait appel à un certain nombre de méthodes statistiques servant à raffiner, simplifier et à rendre le modèle généralisable et facilement interprétable afin d'aboutir à un modèle d'apprentissage capable d'estimer une qualité extrinsèque à partir d'indicateurs intrinsèques.

La dernière étape consiste à valider le modèle d'apprentissage. Dans cette étape, on fait appel à des méthodes statistiques capables d'estimer l'erreur de prédiction ainsi que de tester certaines hypothèses de validation sur un modèle d'apprentissage. Au-delà d'une validation centrée sur un jeu de données, nous proposons d'étudier la transférabilité du modèle sur d'autres jeux de données pris sur d'autres zones d'études et dont les caractéristiques géographiques peuvent varier d'une zone d'étude à une autre afin d'estimer la part de la performance du modèle qui se généralise et qui se maintient sur toutes les zones.

L'ensemble des étapes évoquées ci-dessus constitue une 1ère phase (partie en vert de la figure 4.1) qui vise à fournir un cadre d'estimation de la qualité extrinsèque à partir de la qualité intrinsèque. Cela correspond à la contribution majeure faite dans cette thèse. Par la suite en phase 2 (partie en violet de la meme figure 4.1), nous décrivons une méthode s'appuyant sur le modèle construit lors de la première phase, qui permet une prise de décision pour un utilisateur confronté à une application donnée. L'utilisateur n'aurait pas à se munir d'une base de référence, et pourrait utiliser directement le modèle obtenu à l'issu de la 1ère phase. En fixant un seuil de qualité extrinsèque désiré, l'utilisateur obtiendrait l'ensemble des données dont la qualité est jugée soit inférieure ou supérieure à ce seuil. Quand, la qualité de données est supérieure à ce seuil, nous acceptons ce jeu de données, dans le cas échéant nous le rejetons.

Ainsi, le processus mené par l'utilisateur commence par le calcul d'indicateurs intrinsèques sur son jeu de données. Puis il le soumet au modèle d'apprentissage, qui lui

fournit une réponse sur la qualité extrinsèque de son jeu de données. S'ensuit l'étape de fixation de seuil et la prise de décision sur un objet donné quant à son acceptation pour l'inclure dans l'ensemble des objets pertinents pour son application, en les soumettant éventuellement à des tests de respect des spécifications et des contrôles terrains (par exemple dans le cas où il chercherait à constituer un jeu de référence sur la zone). De ce fait notre méthodologie globale vise à fournir un outil d'évaluation de la qualité d'un jeu de données afin de se prononcer sur la fiabilité dudit jeu de données pour une quelconque application.

4.1.2 Implémentation des indicateurs extrinsèques

4.1.2.1 Démarche de calcul des indicateurs extrinsèques

Dans le but de bien mener une évaluation de la base de données OSM par rapport à sa vérité terrain qui est inaccessible mais que l'on a substitué par une base quasi-référence (BDTOPO), il est souhaitable que les mesures utilisées soient des distances, ce qui permet effectivement de justifier l'utilisation d'un jeu de référence (quoique pour cela seule l'inégalité triangulaire est nécessaire) même si beaucoup de mesures qui ne sont pas des distances ne posent généralement pas de problèmes en général.

En guise de mesures extrinsèques, nous utilisons quatre distances à savoir d'une part la distance radiale et la distance angulaire qui mesurent un écart de forme et la distance surfacique et la distance de Hausdorff qui expriment un écart de position. Par la suite, nous détaillons la démarche d'implémentation de ces métriques en se basant sur la thèse de Bel Hadj Ali (2001b) en y apportant quelques améliorations à travers une étude empirique voire parfois une amélioration purement théorique.

- Distance radiale :
  L'implémentation de la distance radiale se base sur la fonction radiale ou signature polygonale. La signature polygonale dépend du point d'origine et de l'homothétie mais demeure invariante par rotation et par translation. Pour la rendre insensible à l'homothétie, on normalise toutes les valeurs de la fonction radiale par la plus grande valeur. Ainsi, si un polygone est k fois plus grand qu'un autre polygone, leurs signatures polygonales demeurent égales partant du même point d'origine de mesures. La comparaison de la forme de deux bâtiments est donc insensible au facteur d'échelle.

  Par ailleurs, le choix du point de départ de mesures influence la fonction radiale et plu tard le calcul de la distance entre deux signatures radiales. Pour éliminer l'influence du point de départ de mesures, nous procédons un décalage de l'une des deux fonctions par rapport à l'autre de telle sorte à choisir le décalage qui minimise la distance entre les deux fonctions radiales. Ce décalage est un réel τ dans $[0, 1]$ (qui s'exprime en %) comme mentionné dans la définition suivante de d_r.

  Définition (d_r). *Soient A et B deux polygones, on suppose que le point origine des mesures est décalé d'une quantité τ le long du contour du polygone B, la distance $d_r(A, B)$ se traduit par :* $\forall\ s \in [0, 1],\ S_A(s + \tau)$.

  $\forall\ (A, B) \in E^2$,

$$d_r(A,B) = \Big(\min_{\tau \in [0,1]} \int_0^1 \big(S_A(t) - S_B(t+\tau)\big)^2 dt \Big)^{\frac{1}{2}}$$

où $S_A : [0,1] \to \mathbb{R}^+$ (resp. S_B) désigne la signature polygonale de A (resp. B).

Pour que d_r (ou distance angulaire) soit une distance nous devons travailler dans un espace qui soit insensible à un certain nombre de transformations pour lesquelles la distance radiale est invariante. Ces transformations sont en nombre de quatre : rotation, translation, homothétie, et déphasage [1] dans $\mathbb{R}^2$ du point d'origine du polygone. Par soucis de rigueur théorique, on utilise une version continue de déphasage en considérant un batiment comme étant un chemin continu de $\mathbb{R}^2$, c'est-à-dire une fonction $\gamma \in [0,1]$ continue, injective sur $]0,1]$ tel que $\gamma(0) = \gamma(1)$. La représentation graphique du polygone est donc $Im(\gamma)$ et pour une abscisse curviligne s donné, $\gamma(s)$ est le point correspondant du batiment. On précise que l'emprise de bâtiment dans $\mathbb{R}^2$ est invariante par décalage du point d'origine, et qu'a fortiori, n'importe quelle distance entre bâtiments doit avoir la même propriété.

Cependant, pour parvenir à calculer une distance radiale de bonne précision, il est primordial de prêter attention à la construction d'une signature radiale fidèle au polygone. En effet, la fonction radiale étant une fonction continue, il est nécessaire de disposer d'un maximum des points pour pouvoir tracer une courbe assez fidèle à la forme du polygone. Cela nécessite de ne pas considérer seulement les sommets du polygone, mais d'effectuer un sur-échantillonnage sur le contour du polygone. Dans la thèse de Bel Hadj Ali (2001b), en vertu du théorème de Nyquist-Shannon, il est préconisé de sur-échantillonner au moins deux fois les sommets du polygone pour se rapprocher d'une bonne représentation du polygone par une fonction radiale.

Pour illustrer le problème de non échantillonnage, nous examinons l'exemple suivant où on compare la courbe de la signature polygonale d'un polygone de forme carrée en faisant varier le facteur de sur-échantillonnage. Nous remarquons qu'il faut beaucoup plus que 2 fois pour se rapprocher de la signature polygonale correcte (voir figure 4.2). A l'issu de cette expérimentation, nous avons réalisé une étude empirique, recherchant le facteur de sur-échantillonnage optimal réalisant le meilleur compromis entre vitesse de calcul et qualité de la reconstruction auquel il faudrait sur-échantillonner les sommets du polygone.

Ainsi, sur un échantillon de 38500 bâtiments, nous sur-échantillonnons chaque batiment 1000 fois puis on interpole (à défaut d'une fonction analytique) linéairement entre les points pour tracer la courbe de la fonction radiale. On estime que la courbe obtenue représentant la signature radiale, se rapproche très largement de la courbe idéale et est de qualité suffisamment proche de celle d'une fonction analytique pour être utilisée comme référence. Puis on mesure la distance (L_2) entre la fonction analytique et plusieurs signatures polygonales en faisant varier le pas d'échantillonnage, le but étant de voir à combien de fois faut-il sur-échantillonner pour faire converger

1. En pratique le déphasage est une opération discrète consistant à choisir un sommet parmi l'ensemble fini des sommets du batiment. Le déphasage consiste donc en une translation du point d'origine le long de l'abscisse curviligne du batiment.

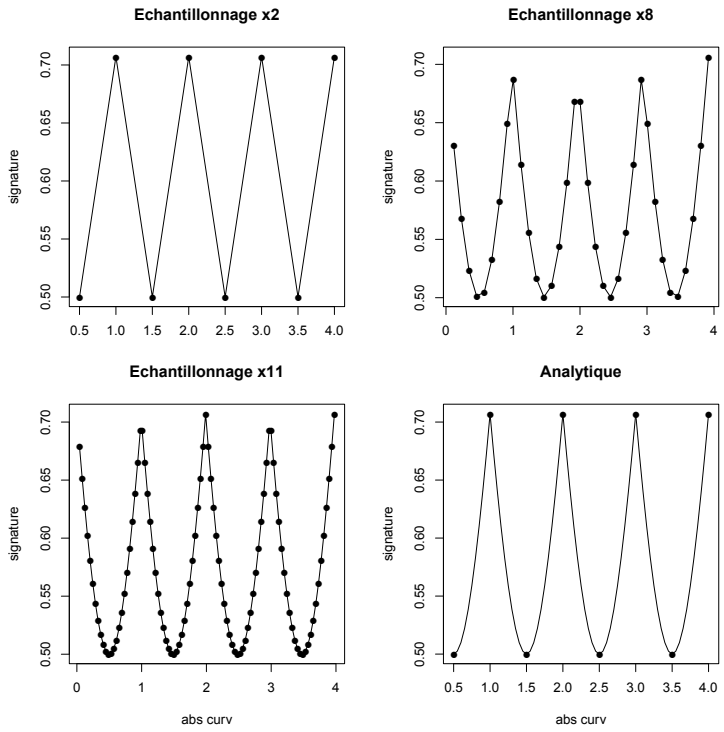


FIGURE 4.2 – Sur-échantillonnage : on représente la signature radiale avec un facteur de sur-échantillonnage variable (2 fois, 8 fois, 11 fois et analytique). La courbe dite analytique correspond à un sur-échantillonnage de 1000 fois qui offre pratiquement la qualité d'une courbe calculée par une formule analytique

l'erreur (la distance calculée) à 0. Le graphique ci-dessous représente l'évolution de l'erreur (Siy) entre signatures polygonales et la fonction analytique en fonction du facteur de sur-échantillonnage (Six). Les lignes en pointillées représentent la bande de confiance à 99%. On observe que pour obtenir une erreur inférieure à 2% sur la distance et en se plaçant sur la borne supérieure de la bande de confiance, il faut sur-échantillonner plus de 11 fois (10.21 fois plus exactement) pour faire converger l'erreur liée au nombre de sur-échantillonnage. La figure 4.3 illustre ces propos.

Dans le cadre de cette thèse, nous avons sur-échantillonné les sommets du polygone 11 fois afin d'atteindre une meilleure précision de construction d'une signature polygonale. Puis, on calcule la signature polygonale sur le polygone échantillonné. Ensuite, on interpole linéairement par défaut, la portion de la courbe entre deux sommets du polygone échantillonné. Enfin, avec le décalage minimal choisi entre deux fonctions polygonales (resp. angulaires), nous parvenons à calculer la distance minimale L_2 des deux fonctions issus des deux polygones échantillonnés en question, et de ce fait, on contourne la variabilité lié au point d'origine et toute autre transformation (rotation, translation, homothétie).

Toutefois, cette démarche de calcul des distances angulaires et radiales peut être améliorée sur deux aspects :

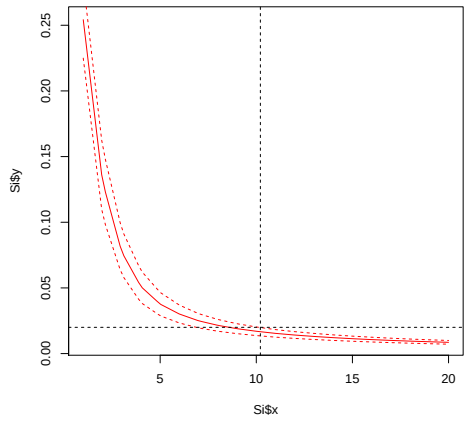


FIGURE 4.3 – Erreur de sur-échantillonnage en fonction du facteur de sur-échantillonnage

1. déterminer l'expression analytique et explicite de la fonction radiale entre deux points consécutifs séparés d'une certaine distance d'abscisse curviligne. Cela permettra de retrouver à l'infini la précision de la fonction de construction de la signature radiale ou angulaire sans être obligé d'effectuer un sur-échantillonnage. NOus pourrions utiliser les sommets initiaux du polygone avec une fonction analytique capable de constituer mécaniquement une infinité de points souhaités au lieu d'effectuer par défaut (à défaut de la fonction analytique), une interpolation linéaire entre deux sommets. De cette façon, la construction de la fonction radiale serait toujours meilleure que celle basée sur un polygone échantillonné tout en éliminant l'erreur engendrée par l'interpolation linéaire.

2. calculer les transformées de Fourier des fonctions radiales ou angulaires : avec notre démarche, pour un ré-échantillonnage de 100 fois, on procède à 100 décalages et 100 calculs de distances. Ce qui donne 10 000 opérations. Or si on passait par le calcul des transformées des fourriers de chacune des deux fonctions (radiales ou angulaires), le nombre d'opérations nécessaires serait bien moindre. Cela réduirait énormément la durée d'exécution du calcul des fonctions pour un nombre conséquent des polygones.

- Distance angulaire : la distance angulaire se base sur la fonction angulaire comme défini dans le chapitre précédent pour mesurer un écart de forme. Cette dernière est aussi sensible au point d'origine mais invariante par translation, rotation et homothétie. Pour remédier à l'influence du point de départ de mesures, on applique le même traitement que celui effectué sur la signature polygonale. Toute la chaine de traitement effectuée sur l'élaboration de la distance radiale, peut aussi s'opérer sur la fonction angulaire afin de déterminer la distance angulaire minimale entre deux polygones suivant le décalage optimal. Après avoir calculer la distance angulaire minimale relative au décalage optimal, il est préférable d'exprimer cette valeur de décalage par un angle $\Delta\theta$, assimilé à une rotation optimale que l'on effectue sur l'un des deux polygones pour obtenir la distance angulaire minimale.

- Distance surfacique : la distance surfacique exprime le taux de chevauchement entre deux polygones. C'est une mesure de la position absolue. Elle vaut le rapport de l'aire de la différence symétrique de deux polygones et de l'aire de leur union, variant entre 0 (complètement superposés ou confondus) et 1 (complètement disjoints).

- Distance de Hausdorff : la distance Hausdorff comme mentionné dans le chapitre 2, constitue aussi une distance de position absolue. Elle mesure l'éloignement entre deux polygones et est atteinte par le maximum d'éloignement réciproque entre les surfaces de deux polygones.

4.1.2.2 Description et illustration des indicateurs extrinsèques

Afin d'illustrer des valeurs typiques d'un batiment vis-à-vis des mesures extrinsèques, et ce pour donner des ordres de grandeurs des distances calculées, nous étudions les distances mesurées progressivement entre un batiment issu la base de référence BDTOPO et d'autres bâtiments issus de la base OSM qui lui sont similaires du point de vue de la forme. En rouge, on a le batiment de la base BDTOPO et en bleu le bâtiments de la base OSM (sauf dans l'image 4.13 sur la distance Hausdorff où un groupe de bâtiments OSM sont représentés en noir). Nous illustrons quelques exemples pour chacune des distances implémentées.

- Distance radiale : on compare la forme d'un batiment BDTOPO tour à tour avec celles de quatre bâtiments de la base OSM, à raison des quatre figures ci-dessous. A chaque fois, on calcule la distance radiale minimale (DR en mètre) avec le décalage optimal ($Phase$ en % sur le graphique).

  Sur les figures 4.4, 4.5, 4.6 et 4.7, sur le 1er graphique (situé à gauche et en haut), on représente les deux polygones des deux bâtiments (un batiment BDTOPO et un batiment OSM) pour lesquels on calculera la distance radiale. Sur le second graphique (situé à gauche au centre), on trace les deux courbes des deux signatures polygonales. Sur le 3ème graphique, on exprime la distance entre les deux signatures pour chaque décalage afin de repérer le décalage qui fournit la plus petite distance radiale. Sur le 4ème graphique (à gauche et en bas) est représenté la distance entre les points des deux signatures en fonction toujours de décalage. On remarquera qu'à chaque fois sur les quatre figures, la distance entre les courbes est minimale pour le décalage optimal repéré sur le 3ème graphique. Sur le graphique 5 (au centre en bas) est tracée la courbe représentant la différence entre les deux signatures polygonales pour le décalage optimal. Et enfin, sur le graphique 6, on représente la distribution de fréquence de la différence entre les deux signatures. Sur ce même graphique, on remarquera aussi que la dispersion de la courbe vaut la distance radiale calculée. Ci-dessous, pour chacune des figures 4.4, 4.5, 4.6 et 4.7, nous indiquons la valeur de la distance radiale ainsi que celle de son décalage minimal qui lui est associé.

  Nous remarquons que le bâtiment issu de la base OSM représenté sur la figure 4.7 semble être le plus proche du batiment BDTOPO. C'est d'ailleurs le seul batiment OSM qui se superpose au batiment de BDTOPO. La bonne similarité de forme amène une bonne similarité de position d'où les deux bâtiments qui se superposent.

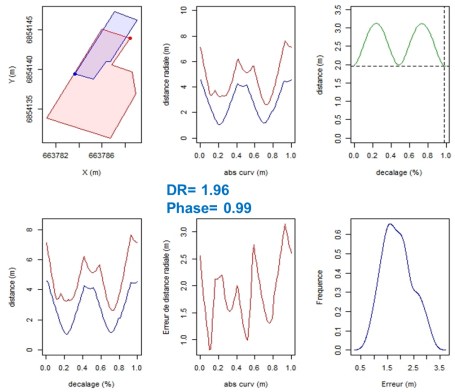


FIGURE 4.4 – Distance radiale : le décalage optimal est de 99% avec une distance radiale égale à 1.96 m

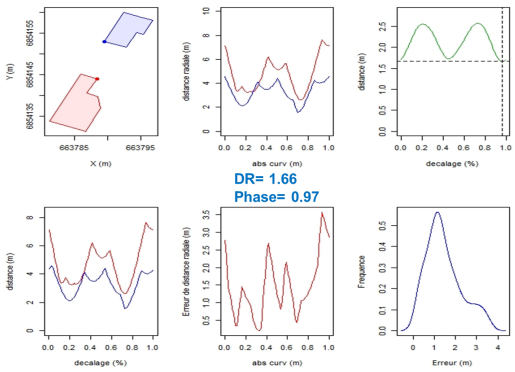


FIGURE 4.5 – Distance radiale : le décalage optimal est de 97% avec une distance radiale égale à 1.66 m

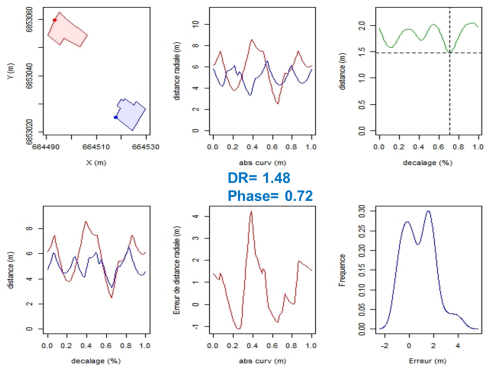


FIGURE 4.6 – Distance radiale : le décalage optimal est de 72% avec une distance radiale égale à 1.48 m

Dès lors on peut de manière empirique, rechercher quelle est la valeur seuil pour la

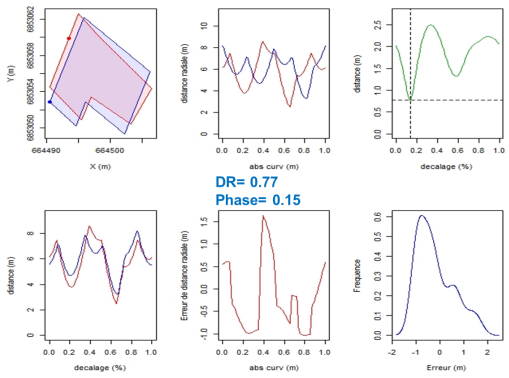


FIGURE 4.7 – Distance radiale : le décalage optimal est de 15% avec une distance radiale égale à 0.77 m

distance radiale au-dessous de laquelle nous croyons fortement que, deux bâtiments sont homologues ou à l'inverse à partir de quelle valeur pour la distance radiale il est très peu probable que deux bâtiments soient homologues. Nous présenterons cette analyse empirique des seuils sur les mesures extrinsèques lors de l'implémentation de l'algorithme d'appariement.

- Distance angulaire : Sur les figures 4.8, 4.9, 4.10 et 4.11, le 1er graphique (situé à gauche d'en haut), représente les deux polygones à comparer. Le second graphique illustre les deux courbes des fonctions angulaires. Sur le 3ème graphique, on représente les différentes distances angulaires en fonction du décalage dans le but de repérer le décalage optimal qui minimise la distance angulaire. Et enfin, le dernier graphique superpose les deux courbes à décalage optimal choisi.

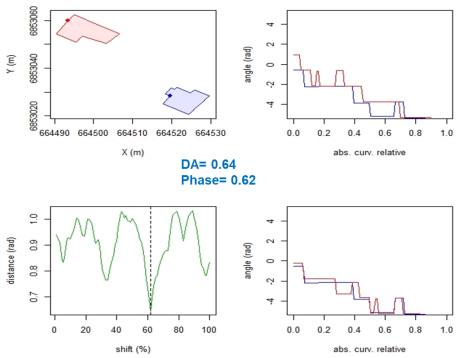


FIGURE 4.8 – Distance angulaire : la rotation optimale est $\Delta\theta = 223.08°$ et la distance angulaire égale à 36.67°

On observe que la meilleure distance angulaire, qui incite à faire croire que les deux bâtiments sont homologues est celle de la figure 4.11. On précise qu'en même, que la distance angulaire est plus sensible que la distance radiale. Un petit changement sur la distance d'angle peut engendrer une modification de forme assez considérable.

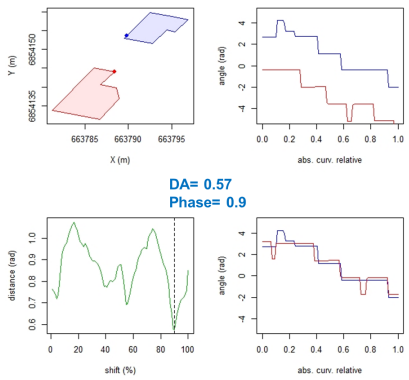


FIGURE 4.9 – Distance angulaire : la rotation optimale est $\Delta\theta = 323.83°$ et la distance angulaire égale à 32.66°

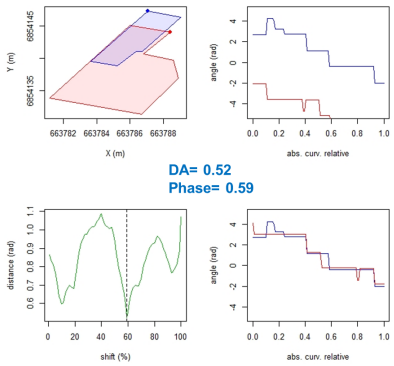


FIGURE 4.10 – Distance angulaire :la rotation optimal $\Delta\theta = 212.32°$ et la distance angulaire égale à 29.79°

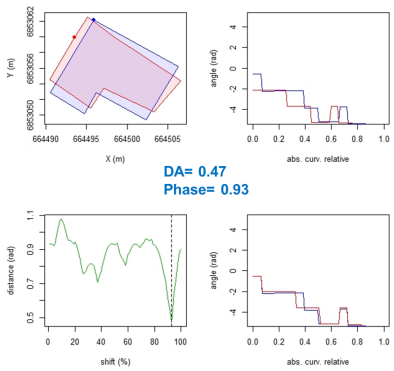


FIGURE 4.11 – Distance angulaire :la rotation optimal $\Delta\theta = 334.63°$ et la distance angulaire égale à 26.92°

- la distance surfacique : sur le graphique à gauche, on représente en noir la surface sur laquelle nous calculons la distance surfacique. Elle correspond en effet à l'aire de la différence symétrique. Sur le graphique à droite sont représenté les deux bâtiments à comparer (voir la figure 4.12.

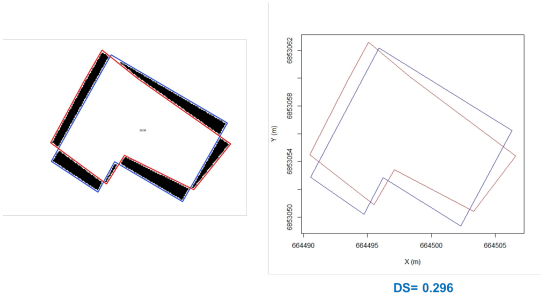


FIGURE 4.12 – Exemple de distance surfacique

- la distance de Hausdorff : nous representons sur la figure 4.13, un batiment BDTOPO entouré par des batiments OSM, candidats à l'appariement. Nous remarquons que le seul batiment susceptible d'etre son objet homologue est bien le seul qui se superpose. La valeur sur la distance de Hausdorff traduit l'loignement maximal réciproque entre un polygone et sommets.

FIGURE 4.13 – Exemple de distance de Hausdorff

4.1.3 Appariement

Dans un contexte de qualification des données de la base OpenStreetMap, et après une phase de d'implémentation des indicateurs extrinsèques, notre travail a consisté à effectuer une correspondance des objets de la base OSM avec les objets d'une autre base de données prise comme une référence. La mise en correspondance des données est la pierre angulaire de ce travail car notre approche se veut précise et robuste en terme de mise en correspondance. Afin de procéder à une comparaison judicieuse, il est impératif de retrouver exactement de part et d'autre dans les deux jeux de données, les objets

géographiques homologues représentant la même réalité sur le terrain quel que soit leur mode de représentation (ponctuel, linéaire ou zonal). Comme nous l'avons vu dans le chapitre 2, l'appariement des données permet réer des liens d'appariement entre les objets homologues de la BDTOPO et d'OSM. L'appariement que nous avons proposé s'effectue en deux étapes : l'appariement de schémas et l'appariement de données.

4.1.3.1 Appariement de schéma

L'appariement de schéma a pour objectif de dresser un premier filtrage de données de manière à identifier les types de données peuvent être appariés ensemble. Il se base sur l'élaboration d'un schéma qui consiste à identifier et à mettre en relation les attributs homologues sur le deux jeux de données. L'appariement de données sera principalement guidé par un tableau de correspondance (schéma). Ce schéma d'appariement a tout son intérêt dans un contexte où l'on veux comparer deux bases de données dont les attributs, les spécifications et les modèles conceptuels diffèrent. En effet, nous remarquons que la structure des *clés* d'OSM permet une description plus détaillée de la sémantique des objets. Ainsi, un *attribut* de la base BDTOPO IGN peut correspondre plusieurs *clés* OSM, en 1ère correspondance, puis un autre attribut en 2ème correspondance. C'est le cas de l'attribut *commercial* de BDTOPO qui peut correspondre dans la base OSM entre autres, aux valeurs *estate agent* (), *bank* (banque), *hairdresser* (laverie), *vehicule-inspection* (garage) en premier lieu, puis à la clé *building* en seconde lieu. Il est possible également que deux attributs de la base BDTOPO puisse correspondre au même attribut dans la base OSM. C'est le cas pour l'attribut *point d'eau* de BDTOPO qui correspond aux valeurs *drinking water* (point d'eau à boire) et *fountain* (fontaine). A l'issu de ces remarques, nous avons mené une étude d'inventaire sur la mise en correspondance de la nature des attributs de la base BDTOPO avec ceux de la base OSM en recherchant d'abord les correspondances premières (c'est-à-dire les correspondances directs) puis les correspondances secondaires. Nous détaillerons ces correspondances dans le chapitre 4. Cela nous a permis préparer l'appariement en confrontant les schémas des deux jeux de données. Il ne reste plus qu'à procéder à l'appariement de données.

4.1.3.2 Appariement des données

L'appariement de données consiste à identifier concrètement les objets homologues à l'aide des critères géométriques et attributaires en se basant sur l'appariement des schémas réalisé auparavant. L'algorithme d'appariement recherche l'objet homologue dans un lot des candidats à l'appariement selon des seuils fixés au préalable.

L'appariement s'appuie sur la notion de cardinalité, de ressemblance et du type de données. Concernant la cardinalité, nous estimons que les deux bases de données ont un même niveau de détail et de contenu proche. Le jeu de données sur lequel nous souhaitons travailler pour déterminer nos indicateurs est l'ensemble des PAI et des bâtiments de la BDTOPO ayant un objet homologue ponctuel ou surfacique dans OSM. De ce fait, nous choisissons une cardinalité 1 à 1 et nous cherchons à apparier chaque objet de la BDTOPO (population de référence) à un objet de la base OSM (population à comparer), la ressemblance étant identifiée à travers un certain nombre d'indicateurs extrinsèques basés sur la forme et la position. Selon le type des données, nous réalisons deux types d'appariement dont l'algorithme est différent. Il s'agit de l'appariement des objets ponctuels et l'apparie-

ment des objets surfaciques. En fonction du type de données (ponctuel ou surfacique) les mesures de ressemblances ne sont pas les mêmes.

4.1.3.3 Performance d'un algorithme d'appariement

Pour pouvoir caractériser la précision spatiale d'un bâtiment OSM (respectivement d'un objet ponctuel), il s'avère nécessaire d'étudier en amont la qualité de l'appariement car nos critères d'appariement nous fournissent les écarts des bâtiments (respectivement des points) OSM avec la base de référence, c'est-à-dire la qualité extrinsèque. La bonne mesure de la qualité extrinsèque dépend du bon appariement des objets. Il est donc nécessaire de valider l'appariement sur un jeu de données test constitué à partir d'un échantillon des *paires d'objets appariés* et des *objets non appariés*.

Durant le processus d'appariement, un algorithme d'appariement peut retrouver le vrai objet homologue, il s'agit donc d'un appariement correct tout comme il peut apparier à l'objet de la base de référence avec un objet non homologue, il s'agira dans ce cas d'un appariement incorrect. Pour pouvoir évaluer la performance d'un algorithme d'appariement , on fait appel à des indicateurs statistiques : *Rappel* et *Précision*.

Le Rappel évalue la qualité d'un algorithme informatique à ne rater d'aucun *vrai objet homologue*. Il exprime le taux de *vrais objets homologues* qui ont été identifié par l'algorithme parmi tous ceux qui existent réellement. Ainsi un très bon rappel signifie que le programme classe très peu d'objets vrais, dans la catégorie des *objets non homologues*.

Par contre, la Précision évalue la qualité d'un algorithme à retrouver que *des vrais objets homologues*. Elle exprime parmi les objets homologues retrouvés par l'algorithme, le taux de *vrais objets homologues*. Elle nous informe sur la capacité du programme à commettre ou pas, des erreurs en identifiant de faux objets homologues (qu'un vrai objet homologue existe ou non).

Les quantités obtenues pour le rappel et la précision nous renseignent sur les caractéristiques de l'appariement. Selon nos exigences, nous pouvons soit valider les résultats d'appariement soit les rejeter.

4.1.4 Implémentation des indicateurs intrinsèques

La 4ème étape de l'approche proposée par la figure 4.1 est de définir les indicateurs intrinsèques. Ces indicateurs pourraient être corrélés avec les mesures extrinsèques sur la précision géométrique et positionnelle. L'objectif est de décrire les objets géographiques à travers des indicateurs intrinsèques caractérisant la qualité de la saisie. Cela nécessite d'analyser les source d'erreurs dans les données géographiques, puis de formaliser les indicateurs intrinsèques comme étant indice d'une éventuelle erreur sur la géométrie et sur la position absolue.

En effet, en étudiant l'origine des imperfections, Batton-Hubert et al. (2019) ont identifié trois sources d'erreurs, les plus courantes qui s'opèrent lors de la saisie de la donnée géographique volontaire. Les auteurs estiment que les erreurs proviennent soit des appareils de mesures, soit d'un manque de connaissance et d'expérience de la personne éditrice de la donnée géographique soit enfin d'un acte volontaire d'origine vandale. Ainsi, sans

toutefois caractériser explicitement ces erreurs, nous formulons nos indicateurs de sorte à pouvoir déceler ces erreurs qui altèrent à la fois la forme et la position d'un polygone. Ces indicateurs pourront donc, nous renseigner sur une mauvaise saisie ou capture des données. Dans le chapitre 5, nous recensons une liste d'indicateurs intrinsèques intrinsèques. Ces indicateurs sont calculés sur les couples d'objets appariés.

4.1.5 Inférence de la qualité extrinsèque à partir de la qualité intrinsèque

A l'issu des étapes d'appariement et de calcul des indicateurs intrinsèques, nous disposons d'un jeu de données d'OSM appariés, où chaque objet géographique d'OSM apparié est labellisé par un ensemble d'indicateurs extrinsèques mesurant les écarts avec son objet géographique homologue provenant de la BDTOPO. Par la suite, nous avons calculé pour chaque objet OSM (apparié), un groupe d'indicateurs intrinsèques qui expriment ses caractéristiques de saisie. Dès lors, nous essayons de rechercher un lien statistique qui relierait les mesures extrinsèques et les indicateurs intrinsèques des bâtiments OSM de sorte à induire une mesure extrinsèque à partir d'un ensembles de mesures d'indicateurs intrinsèques sur un bâtiment donné. On parle de l'inférence de la qualité extrinsèque à partir de la qualité intrinsèque. Les mesures extrinsèques constituent les variables dépendantes tandis que les indicateurs intrinsèques constituent les variables explicatives.

L'inférence statistique que nous proposons dans cette thèse repose sur deux méthodes d'apprentissage. Il s'agit d'une part d'une méthode de régression linéaire multiple couplée avec une méthode LASSO qui est appliquée sur des données géographiques à l'échelle individuelle du bâtiment (échelle micro) ou à l'échelle d'un agrégat de bâtiments (échelle méso). D'autre part, nous proposons d'appliquer une méthode de classification basée sur le principe des forêts aléatoires appliquée à une échelle agrégée (i.e. un agrégat de bâtiments). A travers la méthode de régression, nous cherchons à établir une estimation de la qualité extrinsèque de manière quantitative tandis qu'avec la méthode de Forets Aléatoires, nous souhaitons détecter une mauvaise qualité extrinsèque de manière qualitative à l'aide d'une fonction de classification permettant d'affirmer qu'un objet géographique (i.e. bâtiment) est de est de mauvaise qualité (mal saisi) ou de bonne qualité (bien saisi).

Pour le cas de la régression linéaire, nous examinons l'inférence statistique à l'échelle du bâtiment en premier lieu. Puis dans le but d'améliorer la qualité de la prédiction et ainsi tenir compte en même temps des caractéristiques du bâtiment et de celles de son entourage, nous menons une étude prouvant l'existence d'une auto-corrélation spatiale sur un voisinage. Cela nous incite à regrouper des bâtiments dans le voisinage choisi et à reprendre la régression sur un jeu d'agrégat de bâtiments. Sur cet agrégat, nous calculons et considérons la moyenne des valeurs des bâtiments formant l'agrégat comme d'une part une nouvelle valeur extrinsèque labellisant la nouvelle entité géographique (l'agrégat), et d'autre part une valeur moyenne de celles issus des indicateurs intrinsèques de ce groupe de bâtiments.

En guise de validation de nos modèles d'apprentissages, nous menons des tests statistiques pour vérifier des hypothèses de validation d'un modèle de régression.

4.1.6 Transférabilité de l'inférence

Enfin la cinquième, et dernière étape de notre approche est la transférabilité de l'inférence. Dans l'objectif d'examiner si le modèle issu de l'inférence se généralise, nous testons notre modèle d'inférence de notre zone d'étude sur d'autres zones d'études. On souhaite connaitre la part de la performance de prédiction du modèle qui se transfert (qui se maintient) sur les zones d'études. Au fait, afin d'étudier la performance de prédiction du modèle, nous entrainons le modèle à chaque fois sur une zone d'étude et validons sur le reste des autres zones d'études. A partir de cette étude, nous essayons de caractériser sur chaque zone d'étude, le modèle en matière de performance d'entrainement et de validation afin de retenir le modèle qui aurait la qualité d'être à la fois un bon entraineur mais aussi un bon testeur car ce type de modèle semble avoir la capacité de mieux se généraliser. Sur ce modèle, nous recherchons un seuil de séparation naturelle des données avec lequel on aurait atteint la meilleure performance de prédiction sur les deux autres zones. En dernière section du chapitre sur la transférabilité, nous détaillons deux méthodes de fixation de seuil de qualité utilisées pour deux contextes d'application différents.

4.2 Présentation de la zone d'étude

Notre étude a porté sur une zone extraite du Département *Val-de-Marne (94)* dans la région Ile-de-France. La zone extraite est une fenêtre rectangulaire de $[longueur, largueur] = [6.83km, 4.63km]$ couvrant une surface de $32,84\ km^2$. La zone d'étude recouvre en grande partie au Nord et à l'Ouest deux grandes communes étant parmi les communes les plus peuplées et ayant les plus grandes superficies du département (la Commune de *Créteil*, avec une population de 92265 habitants et une superficie de $11,46\ km^2$, et la commune de *Saint-Maur-des-Fossés* avec une population 75298 *hab* et une superficie de $11,25\ km^2$ en 2018), leur procurant une densité assez forte de l'ordre de 7372 hab/km^2 [2]. A l'Est, elle déborde sur la commune *Chennevières-sur-Marne* relativement 2 fois moins dense [3] (avec une densité environ égale à 3443 hab/km^2) que la partie Nord-Ouest de la zone. Et au Sud, notre zone d'étude s'étend de *Bonneuil-sur-Marne* à *Sucy-en-Brie* et recouvre légèrement la commune de *Ormesson-sur-Marne* sur le Nord, où on passe d'une densité relativement proche de celle à l'Est vers une densité encore plus basse (à Sucy-en-Bry la densité est environ égale à 2586 hab/km^2 [4])(voir la figure 4.14).

A l'issu de cette brève analyse démographique sur les communes, nous pouvons dire que notre zone d'étude se situe dans une zone sous urbaine de population relativement dense. Dans un cade d'apprentissage fondé en partie sur la forme des bâtiments, il est essentiel d'étudier la morphologie urbaine en matière de dimension des bâtiments, de leur densité relative à une sous zone, du type de bâti (résidentiel, commercial...etc.).

Ainsi, nous remarquons dans la figure 4.14 que sur la partie au Nord de la zone, les bâtiments sont davantage très rapprochés tout ayant de petites tailles avec un alignement à la route souvent variable provoquant une disposition et une orientation désordonnée, séparés par des rues venant de toutes les directions. Les formes sont plutôt compactes et leur densité demeure très forte. Dans l'optique de rechercher des regroupements locaux

2. https://fr.wikipedia.org/wiki/Val-de-Marne_Communes_les_plus_peuplees
3. https://fr.wikipedia.org/wiki/Chennevières-sur-Marne
4. https://fr.wikipedia.org/wiki/Sucy-en-Brie

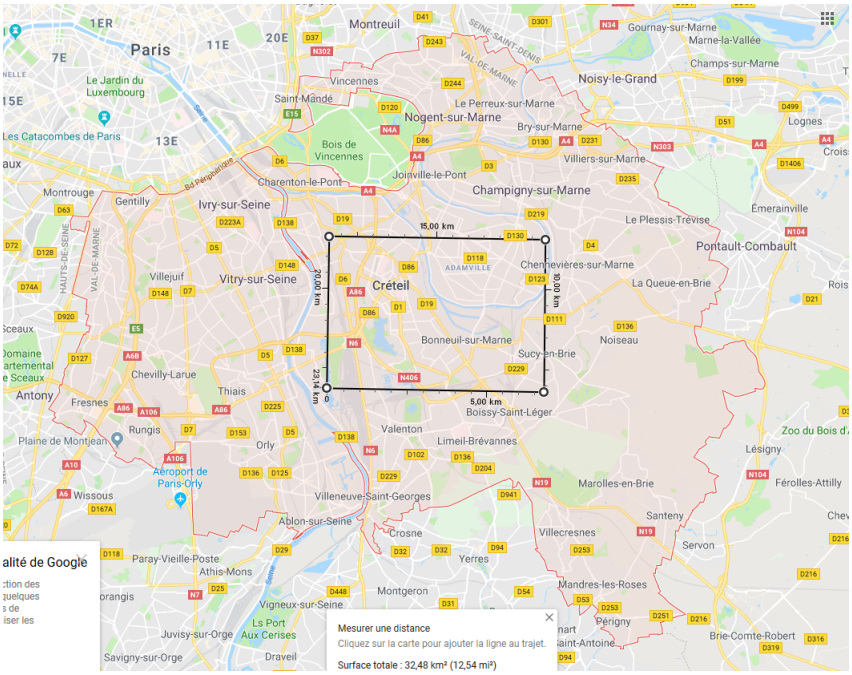


FIGURE 4.14 – Position de la zone d'étude dans le département Val-de-Marne(94)(illustré en rouge). Le rectangle en noir délimite notre zone d'étude.

de bâtiments, on observe très bien apparaitre des ilots urbains séparés par des routes et entourant des zones vides plus ou moins importantes. (voir figure 4.15)

FIGURE 4.15 – représentation du Bâti OSM sur la portion au Nord de la zone d'étude.

Par ailleurs sur la partie Sud-Ouest de la zone d'étude, on observe des bâtiments plus

espacées, de tailles plus grandes, et marqués par des espacements non négligeables, ce qui diminue leur densité. Les bâtiments suivent un alignement faiblement rigoureux. La forme allongée et rectangulaire est plutôt dominante dans cette portion de la zone d'étude. Sur la figure 4.16, nous observons des bâtiments de services publics comme des écoles, des collèges et des hôpitaux.

FIGURE 4.16 – Représentation du Bâti OSMsur la portion au Sud-Ouest sur la zone d'étude..

Enfin, notre analyse morphologique fait ressortir une dernière portion sur la zone d'étude qui se situe au à peu près au Sud-Est. Dans cette partie, on observe des regroupements de bâtiments sur des distances allant jusqu'à 250 m, marqués par une orientation globale progressivement courbée suivant les directions des routes. Sur cette zone, l'intersection des routes fait apparaître des groupes clairs. Les bâtiments demeurent moyennement espacés et de taille moyenne. Toutefois, on trouve certains bâtiments de tailles plus grandes avec une forme allongée vers les extrémités. En somme, La morphologie de cette sous-zone fait apparaître une organisation spatiale plus claire (voir figure 4.17).

Le choix de la zone d'étude répond donc à l'objectif de travailler sur un zone de morphologie diversifiée. Le département de Val-de-Marne a l'avantage de présenter des ruptures nettes entre le rural et l'urbain, ce qui n'est pas le cas de tous les départements. Du Nord au Sud, on passe d'un paysage urbain dense à un paysage rural en passant par une zone fluviale. Le centre fait apparaître une zone industrielle, l'Est est dominé par les bâtiments de type résidentiel, et l'Ouest mélange des bâtiments de type commercial et de type résidentiel dense, sous la forme d'immeubles.

Plus en détail, en s'inspirant de la cartographie fournie par Corine Land Cover (2018), au Nord de la zone d'étude, on trouve un tissu urbain continu dense formé des bâtiments bas avec environ 3 étages tandis qu'à l'Est, on retrouve des bâtiments de type résidentiels pavillonnaires composés d'une habitation avec un petit jardin. Cette zone est marquée

FIGURE 4.17 – Représentation du Bâti OSM sur la portion au Sud-Est sur la zone d'étude..

par un réseau routier régulier. Au Sud, on relève notamment un paysage forestier contenant un parc départemental du domaine de château à des forêts (Sucy en Brie). Du coté de Bonneuil-sur-Marne, nous observons un environnement portuaire renfermant un grand espace industriel et commercial et des chantiers, marqués par de gros bâtiments. Nous observons également un espace vert urbain dans le Sud de Créteil conjointement avec une zone de logements collectifs hauts et denses sous la forme des barres et tours.

Enfin, un dernier élément caractérisant la zone d'étude se base sur la disposition des réseaux routiers et ferroviaires. Nous observons coté Sud à *Sucy-en-Brie* une grosse zone ferroviaire qui semble marquer une rupture nette entre deux typologies de bâtiments. De la même façon au Nord se situent des échangeurs qui forment une limite entre un habitat dense et un habitat résidentiel pavillonnaire. Nous en inférons que ces éléments du réseau peuvent constituer des frontières significatives dans l'optique de regroupements locaux de bâtiments.

Chapitre 5

Appariement de données géographiques

5.1 Appariement des objets géographiques ponctuels

Dans un contexte d'évaluation de la qualité de la base OSM, nous réalisons un processus d'appariement sur deux jeux de données géographiques (un jeu de données de la base OSM et un jeu de données de la base BDTOPO), tout d'abord de type ponctuel, puis de type surfacique. Afin de mieux explorer la complexité du problème lié à l'évaluation de la qualité des données OSM, nous commençons notre processus d'appariement sur les données de type ponctuels nommées *POI* (Points of Interest, en anglais).

5.1.1 Méthode d'appariement des objets ponctuels

5.1.1.1 Appariement de schéma

Comme évoqué précédemment dans la sous section 4.1.3.1, l'appariement de schémas recherche de paires de jeux de données homologues et consiste à identifier les classes homologues, puis les attributs et les valeurs d'attributs homologues des deux bases de données à apparier. Dans la base BDTOPO, les données correspondantes aux POI d'OSM sont organisées dans le thème *Zone d'activité* et concerne les classes : *administratif, bâti, hydrographie, toponyme, culture et loisirs, espace naturel, industriel et commercial, gestion des eaux, orographie, religieux, santé, science et enseignement, transport, et zone habitation*...lister ici la totalité des classes que tu as identifiée, pas seulement un certain nombre. Il s'agit donc d'une correspondance $1 : n$, où $n =$ (remplacer par le nombre exact), c'est à dire que la classe POI d'OSM [1] correspond à xx classes de la BDTopo. Cette première étape correspond donc à l'appariement de classes, même si dans OSM, il n'y a pas la notion de classe.

Une fois les classes homologues identifiées, pour chacune de ces classes, nous avons recherché les attributs de la BDTopo correspondant aux tags d'OSM.

Les tags dans OSM sont organisés sous forme de balise faite d'une combinaison de deux éléments *key ; value*, où la *clé* correspond au nom de l'attribut et *valeur* à sa modalité ou

1. nous utilisons ici la notion de classe par abus de langage car la modélisation de la base de données OSM ne comprends pas la notion de classe

sa valeur. Par exemple, pour signaler qu'un objet géographique est une école, on utilise la clé dite *amenity* avec la valeur *school*. D'autres attributs peuvent être mentionnés, comme par exemple, le tag *building* précisant si l'objet est un batiment (avec la valeur *yes*) ou non.

En effet, nous précisons qu'un objet géographique en OSM peut être cartographier ponctuellement mais le plus souvent en mode surfacique. En mode ponctuel, la nature d'un objet géographique est décrite avec la clé *amenity*[2]. Ce dernier récence les installations utiles et importantes pour les visiteurs et résidents. Ces installations comprennent par exemple, les écoles, les restaurants, les toilettes, les cabines téléphoniques, les pharmacies, les prisons, les toilettes, etc. Pour un objet ponctuel, la saisie se fait à l'aide d'un point placé à une position centrale sur la localisation de l'objet à cartographier. En guise d'exemples, voici quelques valeurs qui peuvent être associées au *tag amenity* : *bar*, *restaurant*, *college library*, *car_rental*, *fuel*, *parking*, *bank*, *atm*, *clinic*, *dentist*, *hospital*, *pharmacy*, *fountain*, *drinking_water*.

En représentation surfacique zonale, si l'objet est composé d'un seul polygone, ce dernier peut toujours être décrit par la clé *amenity =** sauf pour les bâtiments où il faudrait en plus de la clé *amenity =**, spécifier la clé *building = yes*. Par définition, le tag *building* décrit la nature d'un batiment isolé ou un groupe de bâtiments connectés (exemple : les bâtiments d'une résidence universitaire). Bien que l'usage de la clé *building = yes* étant le plus basique, il est nécessaire de définir d'autres valeurs pour la clé *building=** pour donner la nature du bâtiment en question. Aussi, il est à préciser que le type de batiment (ou sa nature) n'est pas nécessairement le même que l'usage (défini avec *building :use=** qui lui est actuellement attribué. Par exemple, un batiment d'hôpital qui est abandonné ou qui a été réaffecté est toujours un *building = hospital*. Pour marquer qu'un hôpital composé d'un seul batiment est actif, on peut utiliser la clé *amenity = hospital* en lui rajoutant le tag *building=yes* ou bien simplement *building = hospital*. Si l'hôpital est un complexe de plusieurs bâtiments et d'autres structures, *amenity = hospital* ne doit pas être appliqué aux bâtiments, mais à l'aire englobant ces derniers.

En fait, pour un objet géographique complexe en mode surfacique, le polygone sur laquelle couvre l'objet géographique est décrit avec le tag *amenity =**, les bâtiments qui s'y trouvent avec le tag *building = school, university, office, house, appartments, hospital, etc* et les équipements et installations associés avec le tag *amenity = parking, toilets, restaurant, shop, etc.* Pour préciser que ces équipements et installations appartiennent à l'objet complexe, il est préférable de leur définir l'attribut *operator* (exemple : un restaurant universitaire avec *amenity = restaurant* en lui rajoutant *amenity = university* et/ou *operator = CROUS / Université Paris-Est*). Lors de la description d'un batiment, l'utilisation du tag *building* peut être générique comme *building = industrial*, *building= commercial*, *building = residential*, etc. La clé peut etre précise comme *building = warehouse*, *building = retail*, *building = appartments*, etc.

Afin de mener à bien la deuxième étape de l'appariement des schémas, à savoir l'appariement des attributs, nous nous intéressons uniquement aux tags qui décrivent la nature des objets géographiques. Ainsi, nous avons identifié que le tag *amenity* d'OSM correspond à l'attribut nature de la BDTopo.

Enfin, la dernière étape de l'appariement de schémas est de définir les correspondances

2. http ://taginfo.openstreetmap.fr/keys/amenity/values

entre les valeurs des attributs *amenity* (OSM) et *nature* (BDTopo). Cette étape, est une étape manuelle et consiste à identifier pour chaque valeur du tag *amenity* d'OSM, la valeur correspondante de l'attribut nature de la BDTopo. Pour cela, nous nous sommes appuyer sur les définitions des valeurs existantes à la fois dans le wiki d'OSM et les spécifications de la BDTopo. Ces correspondances sont dites de *correspondances primaires* résultent de la mise en relation des valeurs aux attributs en se basant sur la signification exacte des valeurs associées à la clé *amenity* (pour OSM) contre celles associés à l'attribut *nature* des PAI de BDTOPO. Le tableau 5.1 montre à titre d'exemples quelques correspondances primaires.

attributs	$Nature(BDTopo)$	$Amenity(OSM)$
	enseignement primaire	*school*
	enseignement superieur	*college*
	parking	*parking*
	point d'eau	*fountain*
	divers public ou administratif	*public*
	etablissement hospitalier	*hospital*

TABLE 5.1 – Exemples illustratifs des correspondances primaires entre les valeurs du tag Amenity(OSM) et les valeurs de l'attribut Nature (BDTopo)

Toutefois, les attributs portés par les objets peuvent être saisies dans l'une des deux jeux de données à apparier de manière plus générique ou plus détaillée. Ainsi les correspondances primaires ne peuvent pas être établies. Ainsi, nous proposons de définir des correspondances approximatives dites de *correspondances secondaires*. Par exemple, les natures de la BDTopo ayant les valeurs *culte islamique*, *culte catholique ou orthodoxe* et *culte israélite* doivent correspondre (i.e. correspondance primaire) respectivement aux valeurs des amenity *mosque*, *church* et *synagoge*. Dans le cas où les correspondances primaires ne peuvent pas être établies, nous souhaitons confronter les PAI BDTOPO à d'autres objets OSM portant la valeur *amenity = place of worship* qui constitue une modalité générique indiquant une nature de type religieuse. Ce type de correspondance constitue une correspondance secondaire comme illustré dans le tableau 5.2.

attributs	$Amenity(OSM)$	$Nature(BDTopo)$
	estate agent	*Divers commercial*
	hairdresser	*Divers commercial*
	research	*Enseignement superieur*
	place of worship	*culte catholique ou orthodoxe*
	retail	*Divers commercial*
	drinking water	*point d'eau*

TABLE 5.2 – Exemples illustratifs des correspondances secondaires entre les valeurs du tag Amenity(OSM) et les valeurs de l'attribut Nature (BDTopo)

Par ailleurs les PAI de la BDTOPO peuvent être confrontés par des objets surfaciques d'OSM. En correspondance primaire, les PAI portant sur des équipements et installations sont confrontés avec des objets surfaciques d'OSM décrits par le tag *amenity = ** tandis les PAI de nature *batiment* sont confrontés aux objets surfaciques d'OSM décrits par le

tag *building* = *.

attributs	$Nature(BDTopo)$	$BuildingouAmenity(OSM)$
	enseignement primaire	$building = school$
	enseignement superieur	$building = college$
	parking	$amenity = parking$
	point d'eau	$amenity = funtain$
	divers public ou administratif	$amenity = public$
	etablissement hospitalier	$building = hospital$

TABLE 5.3 – Exemples illustratifs des correspondances entre les valeurs du tag Amenity et Building(OSM en mode surfacique) et les valeurs de l'attribut Nature (PAI du BDTopo)

Le principe d'établissement des correspondances secondaires lors de la confrontation des PAI BDTOPO avec les objets surfaciques d'OSM s'opère de la même façon que la confrontation se fait entre les PAI BDTOPO et les objets ponctuels d'OSM ou entre PAI BDTPO et les objets surfaciques d'OSM.

L'appariement de schémas consiste ici à trouver les valeurs associées à la clé *amenity* et à la clé *building* qui peuvent correspondre aux différents thèmes des PAI de la BDTOPO énumérés ci-dessus dans la figure 5.3.

L'appariement de schémas est utilisé d'une part pour constituer les jeux de données à apparier (ici c'est l'appariement de classes) et d'autre part pour calculer des distances sémantiques pour l'appariement de données (ici c'est l'appariement des attributs et des valeurs).

5.1.1.2 Appariement des données géographiques

Pour pouvoir apparier les deux jeux de données, nous utilisons un algorithme d'appariement simple que l'on amélioré au fur et à mesure dans le but d'apparier un maximum d'objets mais aussi de réaliser un appariement de bonne qualité. Ces améliorations de l'algorithme d'appariement peuvent consister en l'ajout de critères que nous jugeons utiles, ou en une modification plus profonde du principe de la méthode d'appariement. Nous présentons d'abord les spécificités des bases de données à apparier. Puis, nous détaillons les objets considérés comme candidats et ceux à apparier avec l'ensemble des critères d'appariement pour finir par la démarche de notre méthode d'appariement.

La méthode d'appariement cherche à apparier deux jeu de données, l'un issu de la base de données de référence et l'autre provenant de la base de données OSM. Le jeu de données de référence est celui issu de la BDTOPO tandis que le jeu de données de comparaison est issu d'OSM. L'appariement nécessite un jeu de données dit de référence et un jeu de données dit de comparaison. Pour chaque objet appartenant au jeu de données de référence, la première étape de l'appariement de données consiste à chercher des objets candidats dans le jeu de données de comparaison. L'appariement des PAI BDTOPO avec des objets OSM nécessite d'analyser le jeu de données issu de la base OSM car qu'étant donné la richesse d'OSM on peut se trouver avec beaucoup d'objets candidats OSM proche géométriquement à l'objet à apparier.

En effet, des objets géographiques dans la base de données OSM, tels que les bâtiments, peuvent être représentés soit en mode ponctuel, soit surfacique, soit les deux à la fois. Nous pensons donc que beaucoup d'objets manquants dans OSM en mode ponctuel, existent et sont représentés en mode surfacique. Ainsi, nous ajoutons dans le jeu de données OSM, des objets représentés en mode surfacique de la zone d'étude. Les candidats sont désormais recherchés dans un ensemble d'objets ponctuels et surfaciques dans la base OSM et ceci de manière simultanée. La recherche simultanée dans les deux ensembles de données (ponctuel et surfacique) permet de laisser le choix de mieux retrouver l'objet homologue qu'il s'agisse d'un objet ponctuel ou surfacique en fonction de la distance euclidienne prise comme critère géographique.

De plus, comme déjà mentionné, la base de données OSM contient non seulement des données hétérogènes et et thématiquement plus riches que la base de données topographiques (BDTOPO), mais surtout les modèles conceptuels demeurent très différents. La recherche du plus proche voisin comme objet homologue ne donne pas forcément le bon élément à apparier. En outre, les objets OSM de *nature* différents peuvent porter des *noms* similaires à l'objet à apparier tout en étant près de lui. La recherche de l'objet homologue conjointement par le critère toponymique et le critère géographique pourrait induire une erreur d'appariement. Pour remédier une telle erreur, nous estimons que la solution la plus adaptée à notre cas exige que l'on ajoute à nos critères d'appariement, celui du *type ou nature* des objets. Bien que l'utilisation d'une telle méthode n'autorisant l'appariement qu'entre les objets de même nature résoudrait les erreurs d'appariement (conjointement avec les deux critères précédents), deux objets peuvent candidats peuvent sembler homologues avec l'objet à apparier, selon l'appariement de schéma à cause de la généricité de leurs valeurs liées aux attributs portant sur la nature des objets dans les deux bases (OSM et BDTOPO) comme déjà mentionné dans la partie 5.1.1.1.

A l'issu de ces quelques éléments faisant trait au contexte d'appariement entre un objet PAI BDTOPO et des objets à la fois ponctuels et surfaciques candidats à l'appariement, issus de la base de données OSM, notre méthode utilises trois critères d'appariement à savoir le critère géographique, le critère sémantique et le critère attributaire en définissant des scores pour chaque critère d'appariement.

Ainsi, le critère géographique calcule la distance euclidienne séparant entre un objet à apparier et un autre objet candidat. Si les deux objets à comparer sont ponctuels, la distance euclidienne classique est considérée, tandis que si l'objet candidat d'OSM est un polygone, la distance euclidienne est calculée entre l'objet ponctuel PAI BDTOPO et son projeté sur le segment le plus proche du polygone OSM.

Quant au critère attributaire, portant sur le nom des objets, il compare les deux de chaînes de caractères de manière formelle au moyen de la *distance de Levenshtein*. Pour pouvoir combiner le critère géographique et le critère attributaire, nous devons fixer des scores sur la *distance de Levenshtein*. Si deux objets ont une valeur de *distance de Levenshtein* inférieure à S_L, on attribue à ce candidat comme score, la valeur calculée entre les deux objets. Si par contre l'un de deux objets à comparer, n'a pas de *nom*, on attribue au candidat le score arbitraire égal à *0,8*. Et si enfin, la *distance de Levenshtein* calculée est supérieure à T_L, on attribue au candidat le score arbitraire égal à *0,9*exprimant une grande différence toponymique. Le seuil T_L a été défini de manière empirique.

Enfin, le critère sémantique se base sur le schéma d'appariement. Nous réalisons des correspondances primaires (correspondances exactes pour l'attribut *nature* entre BDTOPO et OSM) ou des correspondances secondaires (correspondances approximatives dues parfois à l'emploi des valeurs génériques ou détaillées dans l'une des deux bases à apparier) quand les correspondances primaires ne sont en mesure d'être établies. Sur la base de la correspondance établie, le critère sémantique réalise une classification des candidats.

L'algorithme d'appariement commence par une étape de sélection des candidats. Il sélectionne les objets ponctuels et surfaciques d'OSM situés à moins d'un seuil T_G1 du point à apparier. Après avoir fait quelques tests, nous parvenons à fixer le seuil pour le critère géographique à T_G2, car au-delà de cette distance, nous pensons que les objets situés ne peuvent plus être des objets homologues à l'objet à apparier. Notre méthode d'appariement opère par la suite sous forme d'un arbre de décision.

En effet, l'arbre de décision effectue pour chaque objet à apparier une classification de ses candidats à l'appariement selon que la nature de l'objet à apparier et la nature de l'objet candidat de la base OSM soient égales (correspondance primaire, classe C1), en correspondance secondaire (correspondance secondaire, classe *C2*), de correspondance non déterminée (classe *C3*, deux objets dont au moins l'une de deux *nature* est inconnue), et enfin différentes (classe *C4*, deux objets dont leur nature est bien connue, mais différente).

Puis s'il existe des candidats attribué à la classe *C1*, on les réattribue de manière ordonnée, dans les sous-classes *C11*, *C21* et *C31*. Pour un candidat donné de la classe *C1*, si la valeur sur la *distance de Levenshtein* est inférieur ou égal à T_L, ce candidat est classé dans la sous-classe *C11*. Dans le cas où, la *distance de Levenshtein* n'a pu être calculée cause d'absence d'attributs, le candidat se voit attribué à la sous-classe *C21*. Mais si la valeur sur la *distance de Levenshtein* est supérieure à T_L (ce qui peut se produire même si l'attribut nature est bien renseigné pour les deux objets) le candidat appartiendra à la sous-classe *C31*. Nous repartissons de la même manière pour les candidats ayant été classé dans les classes *C2* et*C3*, *C4* et *C13*, dans les sous-classes respectives.

Ainsi, on recherche, l'objet homologue de manière progressive et ordonnée en recherchant prioritairement, dans l'ordre, dans les sous-classes suivantes (une classe n'est explorée que si les précédentes ne contiennent aucun candidat) : *C11*, *C21*, *C31*, *C12*, *C22*, *C32*, *C13*, *C23* et *C33*. On précise que nous considérons pas les candidats de *C4* (natures connues et incompatibles = appariement interdit) car nous faisons exclure les liens entre objets de natures incompatibles. A travers cet ordre entre les classes, nous formalisons un indice de confiance dégressif selon le niveau de la classe dans lequel on trouve l'objet homologue, en attribuant un score arbitraire mais dégressif à la classe en question. Puis nous calculons un indice de confiance qui multiplie trois scores : le score sur le critère géographique (distance euclidienne du candidat normalisée par le seuil), le score sur le critère attributaire et enfin le score arbitraire sur le critère sémantique (la classe à laquelle appartient le candidat).

L'algorithme d'appariement basé sur l'arbre de décision permet d'apparier au plus grand nombre d'objets de la base de données de référence tout en diminuant le nombre des liens erronés. Toutefois, la constitution de l'indice de confiance ainsi que le choix des seuils demeurent empiriques et expérimentaux. Nous avons donc cherché à ce que l'indice de confiance soit généré de manière automatique à l'aide d'une fonction de densité prenant

comme paramètres les valeurs des distances, la distance euclidienne (normalisée) et la distance liée au critère attributaire. Pour cela, on doit d'abord procéder à une vérification manuelle des liens sur un échantillon d'objets appariés. Puis pour l'ensemble de liens, la fonction de densité calculera une valeur d'indice de confiance en se basant sur la proportion de liens justes pour les paires ayant des paramètres (classe de correspondance, distance géographique et attributaire) proches. Elle déterminera également les seuils au-delà desquels la probabilité d'apparition de liens erronés devient forte.

Enfin, dans un but de validation des résultats de l'algorithme d'appariement, nous calculons le rappel et la précision de notre méthode, afin d'évaluer la performance du programme d'appariement.

5.1.2 Résultats

5.1.2.1 Évaluation manuelle des liens d'appariement

Afin de valider les résultats d'appariements obtenus, nous constituons d'abord une vérité terrain. Ainsi, nous sélectionnons aléatoirement 60% des liens d'appariement pour procéder à une vérification manuelle (276 liens). Le choix de ce taux se justifie par le fait que le nombre de paires à contrôler n'est pas très important. Durant cette phase, il est essentiel de relever les observations pertinentes afin de mieux analyser le comportement de l'algorithme de l'appariement notamment apporter des explications aux liens erronés. Chaque lien d'appariement sélectionné est qualifié selon trois modalités :

- Appariement sûr : les objets appariés représentent bien la même entité du monde réel ; dans ce cas l'indice de confiance est fixé à 1 ;
- Appariement incertain : nous avons un doute si les deux objets appariés sont homologues ; l'indice de confiance est fixé dans ce cas à 0.5 ;
- Appariement erroné : nous sommes certains que les deux objets appariés ne sont pas homologues ; nous fixons l'indice de confiance à 0 ;

La figure 5.1 illustre un exemple d'appariement sûr où un objet de BDTOPO de nature *enseignement primaire* s'apparie correctement avec son homologue OSM de nature *kindergarten*(garderie en anglais) tandis que sur la figure 5.2 est affiché un objet de BDTOPO de nature *parking* ayant voulu s'apparier à tout prix avec un objet surfacique OSM de nature *parking* se trouvant de l'autre coté des voies ferrés. Sur l'appariement 5.3 nous observons un objet ponctuel BDTOPO qui s'apparie un objet surfacique OSM de nature *building=building*. Comme nous ne sommes pas en mesure confirmer cet appariement, nous le qualifions d'5.3.

5.1.2.2 Analyse des résultats d'appariement préliminaires et ajustements proposés

Quand on observe l'appariement des objets de *nature = enseignement primaire* de la base BDTOPO, on remarque que pratiquement tous les liens sont des *appariements sûrs*. En effet, dans OSM les établissements scolaires sont représentés sous forme de *polygone* en majorité. Si un PAI de *nature = enseignement primaire* de la BDTOPO n'a pas d'homologue de type *point* dans OSM, il y a souvent un homologue de type *polygone* ; la possibilité d'apparier les objets ponctuels de la BDTOPO avec des polygones d'OSM augmente donc le nombre de liens créés. Les objets de type *polygone* sont en général de

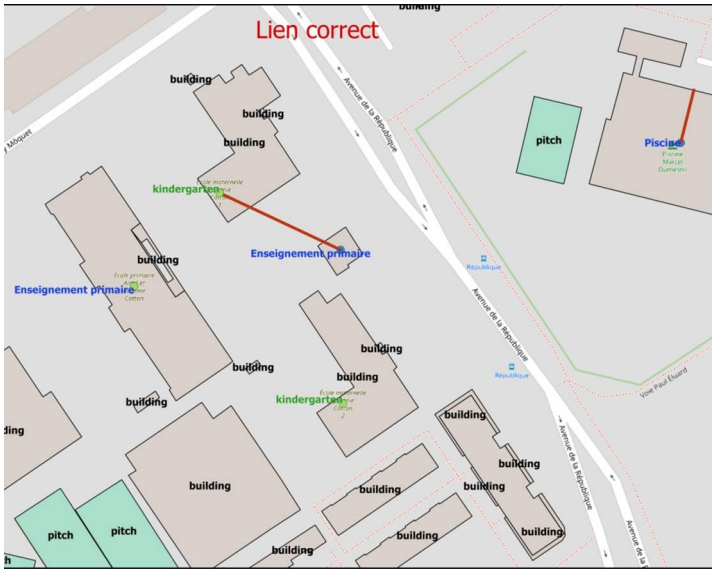


FIGURE 5.1 – Appariement sûr : les objets ponctuels BDTOPO de la base de données de référence (en bleu) sont de nature *Enseignement primaire* tandis que les objets géographiques OSM ponctuels (en vert) de nature *kindergarten* et surfaciques (en noir) de nature *building=building*.

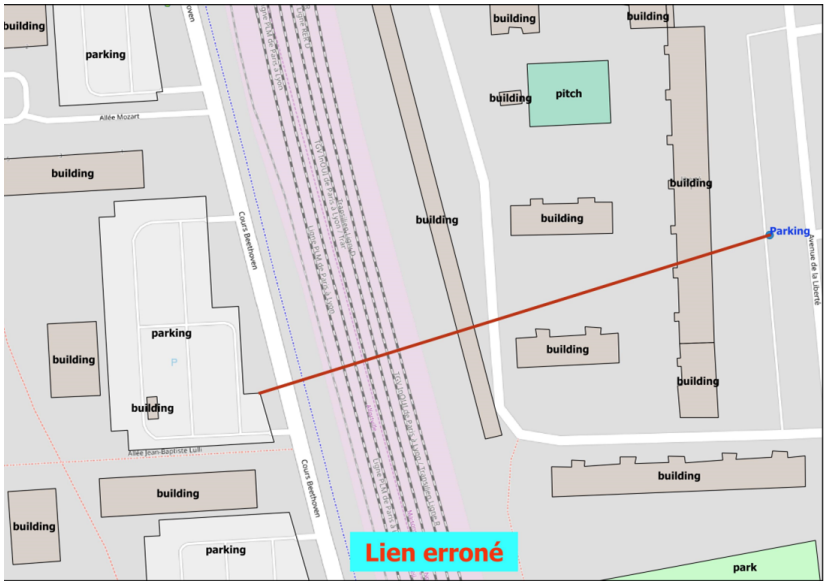


FIGURE 5.2 – Appariement erroné : les objets ponctuels BDTOPO de la base de données de référence (en bleu) sont de nature *Parking* tandis que les objets géographiques OSM surfaciques (en noir) de nature *building=building*

nature *building=building* (cette clé est faussement mis par les contributeurs par défaut) et l'attribut *nom* n'est pas renseigné, mais la vérification nous permet de confirmer l'exactitude de ces liens-là. Quand l'objet OSM existe en mode ponctuel, l'attribut *nature* est souvent renseigné, mais l'attribut *nom* l'est rarement.

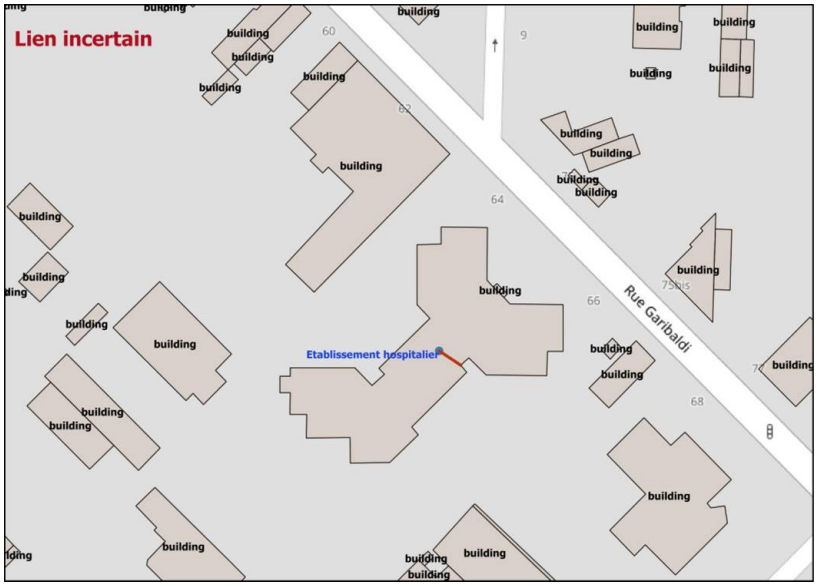


FIGURE 5.3 – Appariement incertain : les objets ponctuels BDTOPO de la base de données de référence (en bleu) sont de nature *Établissement hospitalier* tandis que les objets géographiques OSM surfaciques (en noir) de nature *building=building*

Parfois, quelques erreurs surviennent (appariements erronés) comme le cas de l'appariement d'un objet BDTOPO de nature *enseignement primaire* avec un autre objet OSM de nature *amenity = school* qui n'est pas en réalité son homologue (car le *nom* est différent). L'algorithme cherche à l'apparier d'abord parmi les candidats en *correspondance primaire* alors que parmi les candidats en *correspondance secondaire*, son homologue existe sous forme de *polygone*. Cette erreur d'appariement est due à la priorisation de la correspondance primaire face à la correspondance secondaire d'une part, et à l'imprécision d'autre part. Ces rares cas arrivent au-delà de 90 mètre d'environ. La figure 5.4 illustre ce cas de lien d'appariement. En effet, du fait que les écoles primaires et les écoles secondaires sont désignées par la même nature , il se trouve qu'un objet de la BDTOPO de nature *enseignement primaire* s'apparie avec un objet OSM de type *school* mais qui est concrètement un *collège* alors que le bon objet OSM existe sous forme de polygone mais de nature *building= building*. Ce qui donne un lien erroné lors de la vérification. C'est le cas d'appariement entre le Collège privé *Ozar Hatorah Garçons* (OSM) avec une école primaire (BD TOPO IGN). Ici, la présence de l'attribut *nom* au niveau du collège, a prouvé que ce lien était erroné. Sur la figure 5.4 suivante, en fond de carte, on observe bien le vrai objet homologue étiqueté par le tag *building= building*.

Pour résoudre ces erreurs (le cas d'appariement sur la figure 5.4, il suffit de réduire la distance de sélection des candidats. Ainsi chaque objet de *nature = enseignement primaire* ou *nature = enseignement secondaire* va s'apparier avec le polygone ou l'objet ponctuel le plus proche pourvu qu'il soit le bon élément.

Par ailleurs, les objets de *nature = espace public* (BD TOPO IGN) sont pour la plupart représentés dans OSM en mode *polygone*, et leurs appariements sont souvent validés grâce à la mention de l'attribut *nom* sur la plateforme OSM que l'on utilise pour la vérification. Toutefois, certains objets de *nature = espace public* sont appariés par erreur avec d'autres,

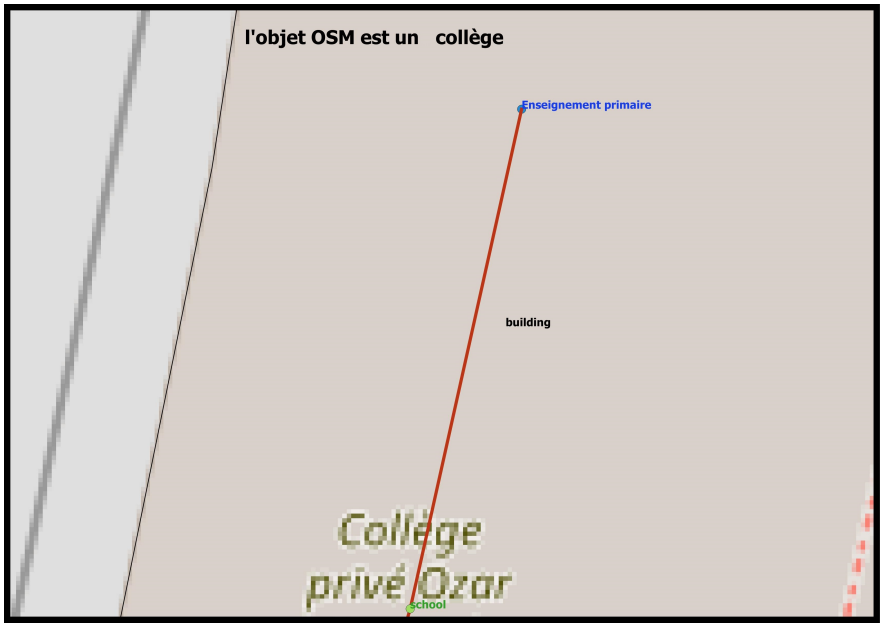


FIGURE 5.4 – Lien d'appariement erroné : objet de la BDTOPO de nature *Enseignement primaire* (en bleu) erroné ici l'objet OSM choisi par l'algorithme d'appariement est l'objet ponctuel OSM de nature *school* au lieu de choisir l'objet surfacique OSM (en vert)

souvent à une distance géographique assez grande car l'algorithme cherche en priorité un objet en correspondance primaire (au-delà de 80m). Ce type d'erreur d'appariement est illustré par le figure 5.5.

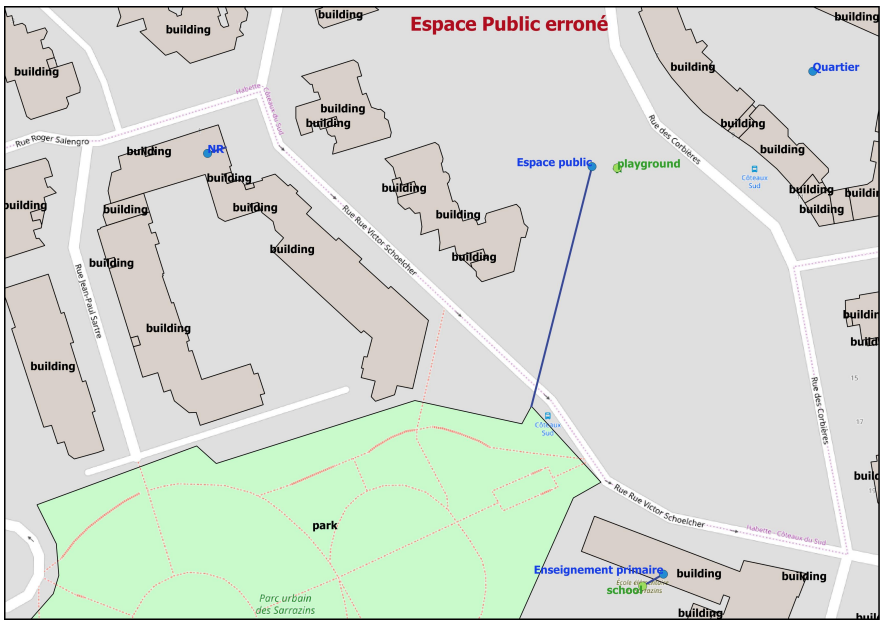


FIGURE 5.5 – Lien d'appariement erroné : objet de la BDTOPO de nature *espace public* (en bleu) qui s'apparie avec un objet géographique surfacique OSM (en vert).

Quant à l'appariement des objets géographiques de type *Établissement hospitalier*, ces derniers s'apparient pour la plupart avec des objets polygonaux de nature *building = building* avec l'attribut *nom* non renseigné et donc aucun moyen de confirmer leur appariement (on leur attribut au score *Appariement incertain*). Cependant, d'une manière générale, quand les objets polygones (dans OSM) ne portent pas leur toponyme, on recourt à la forme du polygone et si l'objet (BDTOPO) se trouve dans le polygone, on tend à le valider surtout que la plupart de ces polygones se situent à des faibles distances de l'objet de BDTOPO. Nous avons un exemple de cas sur la figure 5.6.

Concernant les PAI de *nature = Point d'eau* de la BDTOPO, on peut les apparier avec des objets de nature *amenity = fountain* ou *amenity = drinking water*. En principe, le type *fountain* est plus généralisé que celui de *drinking water*. Or, dans la *correspondance primaire* on autorise uniquement à ce que *Point d'eau* s'apparie avec *drinking water*. Cela amène parfois des erreurs d'appariement d'un *Point d'eau* avec un *drinking water* situé à *60 m* alors qu'il y avait un vrai équivalent (un objet de type *fountain*) situé à quelques mètres. Désormais on autorisera les deux correspondances au même niveau. C'est le cas de la figure 5.6 suivante, où l'algorithme d'appariement a cherché plus loin un objet OSM de nature drinking water alors que le vrai objet homologue OSM existe en nature *fountain* tout près de l'objet BDTOPO de nature *Point d'eau*.

FIGURE 5.6 – Lien d'appariement erroné sur un point d'eau

5.1.2.3 Évaluation quantitative des résultats d'appariement

A l'issue de nos observations sur les liens erronés d'appariement, nous avons pu corrigé le processus d'appariement de l'algorithme afin d'être confronter seulement aux qualités intrinsèques de l'algorithme d'appariement. Plus explicitement, l'analyse préliminaire d'appariement de données nous a permit d'améliorer l'appariement de schémas, et que cela permet de mieux évaluer la qualité de ce processus d'appariement.

A l'issue du lancement de l'algorithme d'appariement, nous observons que parmi les **681** objets ponctuels de la BDTopo à apparier, **462** objets ont été appariés avec des objets d'OSM : **150** objets de type polygones et 312 objets de type ponctuels. Les objets non appariés ont pratiquement tous des candidats dans la base OSM, car les seuils choisis sont assez larges (distance géographique maximale = **170 m**, seuil de Levenshtein = **0.76**). Le lien d'appariement entre deux objets homologues est modélisé par un objet géographique de type linéaire. Pour deux objets homologues de type ponctuels, le lien d'appariement relie les deux objets homologues tandis que pour deux objets homologues de type ponctuel et surfacique, le lien d'appariement relie l'objet ponctuel et le segment le plus proche de son homologue de type surfacique. En plus de sa géométrie, chaque lien d'appariement est caractérisé par des attributs tels que la distance géographique, la distance de Levenshtein.

Enfin, dans un but de validation des résultats de l'algorithme d'appariement, nous calculons le rappel et la précision de notre méthode, afin d'évaluer la performance du programme d'appariement.

5.1.2.4 Performance du programme d'appariement

Pour pouvoir définir les modalités de calcul des indicateurs de performance, on définie cinq possibilités d'appariement :

1. Il existe un équivalent vrai et il est trouvé par l'algorithme : (I)
2. Il existe un équivalent vrai mais un faux élément est trouvé : (II)
3. Il existe un équivalent vrai mais il n'a pas été retrouvé par l'algorithme : (III)
4. Il n'y a pas d'équivalent vrai mais 1 lien faux est trouvé : (IV)
5. Il n'y a pas d'équivalent vrai et rien n'a été trouvé : (V)

On définit pour les objets appariés le Rappel comme suit : $R = (I)/((I)+(II)+(III))$. C'est la proportion de liens trouvés parmi les liens corrects.

Et la Précision pour les objets appariés se calcule comme suit : $P = (I)/((I) + (II) + (IV))$. C'est la proportion de liens corrects parmi les liens trouvés

Pour les objets non-appariés, le Rappel se calcule comme suit : $R = (V)/((IV) + (V))$. Pour les objets non-appariés, la Précision se calcule comme suit : $P = (V)/((III) + (V))$.

A l'issue de notre vérification manuelle des liens d'appariement, nous avons pu classer les différents liens d'appariement suivant ces 5 possibilités. Ainsi, nous avons pu obtenir des indicateurs de performance :

- Pour les objets appariés : P = **96.5%** et R = **94.46%**
- Pour les objets non-appariés P = **39.28%** et R = **78.57%**

Cela signifie que ce programme commet **3 à 4%** d'erreur avec encore **5%** d'omission des bons éléments.

5.1.2.5 Indice de confiance

Nous nous sommes intéressé à déterminer nos seuils d 'appariement de manière automatique et relative à nos données. Pour cela on fait appel à une méthode non-paramétrique

d'estimation de la densité de probabilité d'une variable aléatoire. Elle se base sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support. Pour notre cas, nous passons du nuage de points (corrects et erronés) à la probabilité qu'un lien soit vrai avec une fonction de densité de noyau. A partir de ces deux groupes de probabilités, et à l'aide de la fonction de densité de noyau, on parvient à calcul la probabilité qu'un lien soit vrai sachant un couple de distance géographique et distance de Levenshtein. Cette probabilité exprime l'indice de confiance que nous accordons aux liens d'appariement. Elle nous permet aussi de savoir au-delà de quelle distance géographique et de quelle distance de Levenshtein, nous estimons que l'algorithme crée plus de liens erronés que de liens corrects.

On note L la distance de Levenshtein et G la distance géométrique. L'idée est de calculer la probabilité qu'un appariement soit correct (points verts sur la figure) sachant les valeurs L et G.
Si on note $A = 1$ l'évènement *appariement correct* et $A = 0$ l'évènement *appariement incorrect*, on cherche à estimer :
$P(A = 1|L, G)$
D'après la loi de Bayes, on peut écrire :

$$P(A = 1|L, G) = P(L, G|A = 1) * P(A = 1)/P(L, G)$$

Par ailleurs :

$$P(L, G) = P(L, G|A = 0) * P(A = 0) + P(L, G|A = 1) * P(A = 1)$$

En combinant les deux relations ci-dessus, on trouve :

$$P(A = 1|L, G) = P(L, G|A = 1)*P(A = 1)/(P(L, G|A = 0)*P(A = 0)+P(L, G|A = 1)*P(A = 1))$$

avec :

- $P(A = 1)$ et $P(A = 0)$ les probabilités a priori, qu'on peut estimer à partir du ratio de points verts et rouges sur le nombre total de points ;
- $P(L, G|A = 1)$ que l'on peut calculer à partir de l'estimation par noyau dans l'espace (L,G) de la densité du semis de points verts $(A = 1)$;
- $P(L, G|A = 0)$ que l'on peut calculer à partir de l'estimation par noyau dans l'espace (L,G) de la densité du semis de points rouges $(A = 0)$.

  La figure 5.7 représente à gauche les densités $P(L, G|A = 1)$ et $P(L, G|A = 0)$ estimées par noyaux et représentées par des courbes de niveau. On observe que la probabilité d'apparition des appariements corrects est forte (les points verts) pour une distance de Levenshtein inférieure à 0,76, et pour L=0,8 avec une distance euclidienne inférieure à 70m. Elle persiste faiblement jusqu'à 100 m. On observe que les appariements erronés se concentrent à l'extrême droite de 100 m. A droite, on représente la probabilité d'appariement correct en fonction de (L, G). On constate que cette probabilité d'appariement correct diminue fortement entre 80 et 90m. A partir de ce graphe, nous estimons que le seuil de la distance géographique optimal est de 90 m.

L'appariement des points nous a servi dans un premier temps à appréhender le problème de l'incomplétude des données OSM en représentation ponctuelle. Ce qui induit des erreurs dans l'appariement avec les données de la base de référence. Nous pensons également

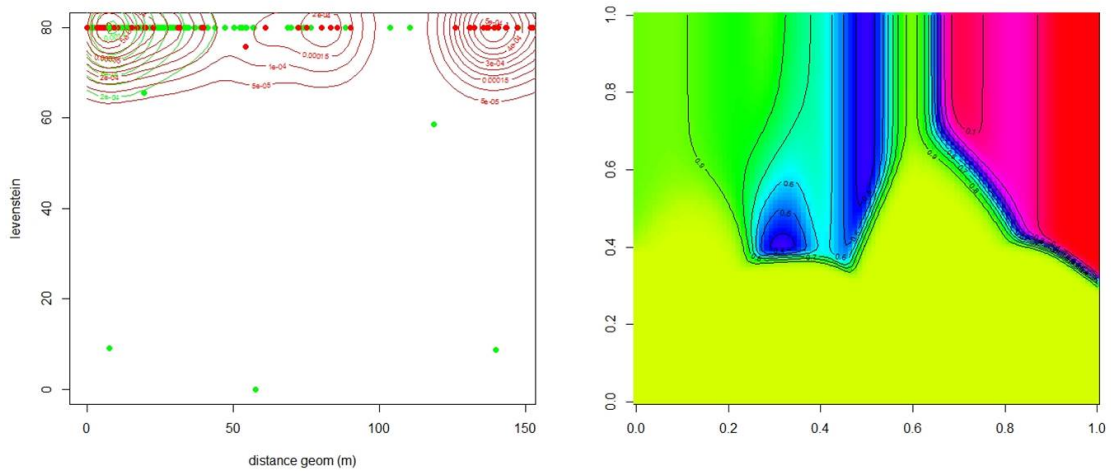


FIGURE 5.7 – Indice de confiance à partir de fonction de Kernel (noyau)

qu'une ambiguïté liée aux attributs des données OSM peut engendrer aussi des appariements erronés car deux objets OSM peuvent porter le même attribut. La redondance issue du fait qu'une entité géographique peut exister à la fois en mode ponctuel et mode surfacique, peut amener à des appariements variables répondants au plus proche voisin. Or on rappelle que l'appariement constitue la base de notre approche. Il doit être robuste et très fiable. C'est pourquoi nous avons songé à privilégier dans la suite le travail sur des bâtiments représentés sous forme surfacique dans la BDTOPO et dans OSM. Ce choix des données surfaciques est conforté par le fait que la plupart des entités géographiques dans OSM existent en représentation zonale. Pour dépasser l'hétérogénéité géographique, thématique et sémantique des points, on s'appuiera sur la forme et la position des entités surfaciques en établissant de nouveaux critères géométriques et positionnels dans un algorithme multi-critère basé sur les fonctions de croyances. Il est important que nous disposions d'un appariement solide et rigoureux pour que la suite de notre étude puisse être fiable.

5.2 Appariement des objets surfaciques

5.2.1 Méthode d'appariement des objets surfaciques proposée

5.2.1.1 Introduction à l'appariement basé sur la théorie de croyances

La mise en correspondance des données est la pierre angulaire de ce travail car notre approche se veut précis et robuste en termes de mise en correspondance. Nous avons choisi de fonder notre algorithme d'appariement sur la théorie des fonctions de croyance. L'appariement basé sur la théorie des fonctions de croyance est une solution flexible, qui permet de prendre en compte les différents critères d'appariement que nous avons construits pour les entités surfaciques, sans avoir à calculer une fonction de score globale qui peut conduire à un comportement instable de l'algorithme. Comme défini dans Olteanu-Raimond et al. (2015), cette méthode est adaptée car elle permet de prendre en compte les cas où les données pourraient être manquantes, incertaines ou imprécises. Dans

le manque de données, l'algorithme prend en compte par exemple le manque de l'information thématique pour un objet donné, le critère thématique ne peut pas être utilisé car pas de mesure ; l'algorithme s'adapte et prend une décision en s'appuyant sur les autres critères ;on également le cas où une entité du monde réel n'est pas représentée dans un jeu de données, l'algorithme s'adapte en permettant la solution *non-apparié*.

L'appariement basé sur la théorie des croyances, permet de fusionner les connaissances issues de plusieurs critères d'appariement pour finalement décider une solution parmi trois solutions possibles : *apparié*, *non-apparié* et *indécis*.

Dans cette thèse, nous étendons la méthode proposée dans Olteanu-Raimond et al. (2015) en modélisant et en définissant des critères capables de mesurer les similarités entre les polygones. Pour la correspondance des polygones, nous proposons d'utiliser deux critères géométriques et deux critères de position. Pour les critères géométriques, nous considérons la distance angulaire et la distance radiale, tandis que pour les critères de position, nous avons choisi la distance surfacique et la distance de Hausdorff. Ces distances sont calculées entre un objet géographique d'un jeu de données de référence issu de la BD-Topo et d'un groupe d'objets géographiques candidats d'un jeu de données issu d'OSM. À ce stade de la mise en correspondance des données, la valeur ajoutée de ce travail par rapport aux travaux précédents, comme ceux dans Olteanu-Raimond et al. (2015), apparaît à deux niveaux. Premièrement, nous modélisons des connaissances permettant la définition des critères d'appariement prenant en compte les caractéristiques géométriques et de position des objets surfaciques. Deuxièmement, nous mettons en œuvre la méthode d'appariement basée sur les nouveaux critères et appliquée sur des jeux de données ayant une géométrie surfacique d'une part et la définition des seuils adaptés à ce types d'objets géographiques, d'autre part.

5.2.1.2 Appariement des données géographiques surfaciques

Comme proposée par Olteanu-Raimond et al. (2015), la méthode d'appariement est composé de cinq étapes. La première étape est appelée la sélection des candidats. Pour chaque objet géographique du jeu de données de référence (featureRef), nous recherchons des candidats dans l'ensemble du jeu de données d'OSM ($candidatOSM_i$). Pour optimiser le temps de recherche de candidats, nous utilisons des index spatiaux. Sans l'indexation, chaque recherche d'objet géographique nécessite d'accéder séquentiellement à tous les objets géographiques de la base de données. L'indexation nous a permis d'organiser les objets géographiques d'abord dans un espace divisé en cellules carrées (de 30 m de coté). Puis nous identifions la cellule de l'objet géographique du jeu de référence, et enfin nous recherchons les candidats dans les 9 cellules voisines de telle sorte que ceux se trouvant à 9 cellules autour de la cellule de l'objet géographique du jeu de référence, soient seulement sélectionnés. Les 9 cellules constituent chacun un carré de 30 m de côté, ce qui correspond au seuil fixé pour la sélection de candidats. Avec ce seuil, nous pensons que l'objet homologue est systématiquement sélectionné parmi les candidats. Cela a donc accéléré la phase de recherche de candidats.

La deuxième étape consiste à initialiser les masses de croyances. Pour chaque couple d'objets géographiques (featureRef, candidatOSM) nous comparons les connaissances issues des différents critères. Pour chaque critère d'appariement, trois fonctions de croyance sont définies pour chacune des trois hypothèses, à savoir l'hypothèse le candidat est l'objet homologue ($appCi$), le candidat n'est pas l'objet homologue ($-appCi$) et l'hypothèse d'ignorance je ne sais pas si le candidat est l'objet homologue (Θ). Ces trois fonctions, ex-

primant des poids accordés à chaque hypothèse, matérialise la croyance que l'on accorde à un candidat a priori (sans connaître les poids des autres candidats). La somme des masses de croyances vaut 1 sur l'ensemble de trois hypothèses. Le tableau 5.4 suivant résume l'ensemble des seuils impliqués dans les équations des fonctions de croyance. Une bonne définition des fonctions de croyance permet de configurer l'algorithme d'appariement des données de manière à le faire converger vers la décision la plus plausible ou la plus crédible en minimisant les conflits qui surviennent lorsque deux critères soutiennent simultanément deux candidats distincts. Ainsi, afin de définir des seuils pour les fonctions de croyances, nous avons dû mener une étude empirique en observant la distribution des valeurs des mesures (c'est-à-dire les valeurs de distances surfacique, angulaire , etc.).

L'analyse des valeurs de mesures des similarités, nous a permis d'observer une grande similitude des valeurs entre *featureRef* et chacun des candidats *candidatOSM* pour les critères géométriques. Nous décrivons par la suite la modélisation des connaissances pour les critères géométriques.

Critère géométrique basé sur la distance angulaire
. Le critère géométrique basé sur la distance angulaire repose sur la comparaison des formes de deux polygones. Pour un objet géographique *featureRef* donné, nous avons observé que plusieurs candidats peuvent avoir des distance angulaires faibles et similaires ou proches. Si nous associons un poids (masse de croyance) fort à l'hypothèse *appCi* pour des faibles valeurs de distance angulaire, nous autorisons à ce que le critère sur la distance angulaire soutienne plusieurs candidats, ce qui pourrait conduire à un conflit fort. Pour réduire le conflit, le résultat (masse) sur le critère géométrique basé sur la distance angulaire doit être corroboré avec les valeurs (masses) tirées des critères de position. Ce qui nécessite d'attribuer une masse de croyance assez forte à l'hypothèse d'ignorance, malgré que les valeurs sur la distance angulaire sont fiables. Nous laissons le choix aux critères de position, de prendre la décision sur le candidat le plus probable parmi l'ensemble des candidats. Au fur et à mesure que les valeurs sur la distance angulaire augmentent, l'hypothèse d'ignorance se voit renforcée autant que l'hypothèse de *non-apparié* augmente de façon moindre. C'est pourquoi sur la figure 5.8 sur la distance angulaire, nous attribuons une masse de croyance égale équitable (égale à 0.5) à l'hypothèse *appCi* et l'hypothèse Θ pour des valeurs de distance angulaire très faibles. Et dès lors que la valeur sur la distance angulaire croit légèrement jusqu'à une valeur seuil (T_{1a}), la masse de croyance sur *appCi* décroit rapidement vers 0 faisant douter d'avantage du candidat voire à croire qu'il n'est plus l'objet homologue. Nous avons donc besoin de déterminer le seuil sur lequel opère le changement de l'allure des courbes des fonctions de croyance. Ce seuil correspond à la valeur sur la distance angulaire au-delà du-quelle nous doutons fort que l'objet géographique *featureRef* et l'objet géographique *candidatOSM* soient homologues. Nous avons mené une étude empirique pour déterminer la valeur sur le seuil T_{1a}.

Critère géométrique basé sur la distance radiale
. Tout comme le critère sur la distance angulaire, le critère géométrique basé sur la distance radiale mesure la ressemblance ou la dissemblance de formes entre deux polygones. Les valeurs sur la distance radiale sont aussi faibles et proches pour un groupe de candidats. Il est difficile de parvenir à distinguer parmi ces candidats l'objet géographique homologue à *featureRef* en utilisant uniquement le critère géométrique basé sur la distance radiale. Pour initialiser les masses de croyance sur les trois hypothèses, nous adoptons la même démarche que celle utilisé pour initialiser les masses de croyance du critère basé sur la dis-

tance angulaire. Nous reprenons les mêmes fonctions de croyance et les mêmes masses de croyance que celles dressées pour le critère sur la distance angulaire. La différence notable s'opère au niveau du choix du seuil sur la distance radiale à partir du-quelle, l'hypothèse $appCi$ tend à s'annuler. Dans une étude empirique, nous avons recherché la valeur du seuil à partir du-quelle nous croyons de plus en plus mais faiblement que le candidat n'est plus l'objet homologue avec une grande prépondérance de l'hypothèse d'ignorance. Les fonctions de croyance pour le critère basé sur la distance angulaire représentées dans la figure 5.8 (en haut) sont utilisées pour décrire l'initialisation des masses de croyance pour le critère basé sur la distance radiale.

Critère de position basé sur la distance Hausdorff

Le critère de position basé sur la distance Hausdorff mesure l'éloignement entre de deux objets géographiques. L'initialisation des masses de croyances pour les trois hypothèses $appCi$, $-appCi$ et Θ est représenté en figure 5.8 (en bas). Notre hypothèse est que plus deux objets se superposent, plus le critère croit que les deux objets soient appariés et plus deux objets sont disjointes, plus il croit que les deux objets ne doivent pas être apparié. Les masses de croyances pour le premier cas de figure, varient de $E_{ds} = 0.01$ à 1 au fur et à mesure que la valeur sur la distance de Hausdorff décroit de T_{2h} à T_{1h} (voir figure 5.8 et table 5.4) pour l'hypothèse $appCi$. Pour le second cas de figure (les objets disjoints), les masses de croyances faisant en sorte que le critère soutient de plus en plus à l'hypothèse de $-appCi$, passent de 0 à 0.6. Aussi, de T_{1h} à T_{2h}, l'hypothèse d'ignorance prend d'importance en passant de 0.2 à 0.4 comme illustré sur la figure 5.8.

Critère de position basé sur la distance surfacique

Ce critère mesure l'éloignement de deux objets géographiques et s'appuie sur la distance surfacique qui varie entre 0 (featureRef et candidatOSM se superpose entièrement) et 1 (featureRef et candidatOSM ne s'intersecte pas). L'initiation des masses de croyances pour ce critère suit la même logique que le critère basé sur la distance de Hausdorff. Seuls les seuils sur la mesure de similarité ont changé. Nous fixons pour les seuils sur la distance surfacique 0.5 et 0.6 (5.4).

A travers ces remarques, nous parvenons à dresser une allure sur chacune des fonctions de croyance sur les critères géométriques. Pour chaque indicateur, nous définissons un critère d'appariement. La figure 5.8 illustre les fonctions de croyance pour les critères d'appariement basés sur les distances angulaires (en haut) et de Hausdorff (en bas). Notez que le critère d'appariement basé sur la distance radiale suit des fonctions de croyance similaires à la distance angulaire, à l'exception des seuils sur les masses de croyances qui sont légèrement différent (voir tableau 5.4). Il en est de même pour la distance de Hausdorff et la distance surfacique.

La troisième étape de l'algorithme de données est la fusion des critères pour chaque couple (featureRef, candidatOSM). Elle consiste à fusionner les masses de croyances associées aux trois hypothèses de chacun des quatre critères.

La quatrième étape de l'algorithme consiste à fusionner les candidats. Plus précisément, les masses de croyances fusionnées par candidat sont maintenant fusionner en prenant en compte tous les candidats ensemble. Nous rappelons que la probabilité pignistique est une fonction de probabilité intervenant au niveau de la prise de décision (Smets et Kennes,

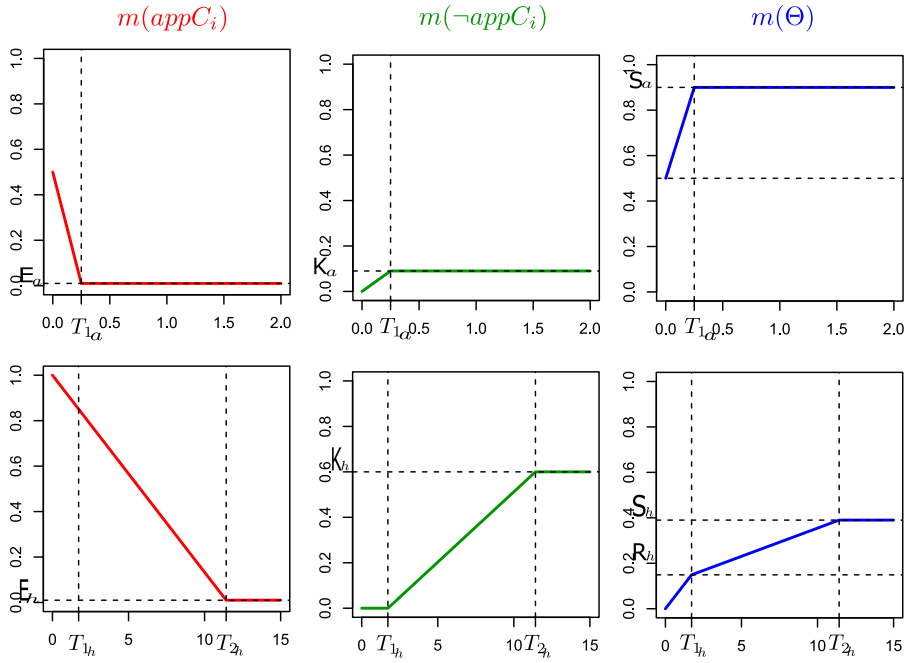


FIGURE 5.8 – Fonctions de croyances : établissement des masses de croyances pour chaque hypothèse ($appCi$, $\neg appCi$, and Θ), pour la distance angulaire da (en haut) et la distance de Haussdorff dh (en bas) avec leurs valeurs de masses de croyances (E_a, K_a,S_a for da, and E_h and K_h,S_h,R_h for dh) en fonction de leurs seuils (T_1 et T_2)

settings	T_1	T_2	E	K	S
da	0.25	$--$	0.01	$0.1 - T_1$	0.9
dr	0.7	$--$	0.01	$0.1 - T_1$	0.9
dh	1.72	11.42	0.01	$1 - E - S$	0.6
ds	0.5	0.6	0.01	$1 - E - S$	0.6

TABLE 5.4 – Settings of belief functions of angular distance(da), radial distance (dr), haussdorf distance (dh) and surface distance (ds) as the criteria of the data matching

1994; Smets, 1998; Olteanu-Raimond, 2008).

La cinquième étape, quant à elle, consiste à calculer la probabilité pignistique pour chaque hypothèse issue de la fusion des candidats.

Enfin la dernière étape est l'étape de décision. Cette étape est différente de celle proposée par Olteanu-Raimond et al. (2015). En effet, au lieu de calculer la différence entre les deux premières valeurs maximales des probabilités pignistiques, nous proposons de calculer le rapport entre le premier maximum et le deuxième maximum des probabilités pignistiques.

Pour justifier ce choix, nous prenons un exemple d'appariement correct mais pas trouvé (à cause de la possibilité d'indécision). Ce cas là correspond à la base un résultat du cas des *objets appariés*, mais en réalité l'algorithme n'a pas apparié en raison d'un conflit entre les deux probabilités pignistiques de deux premiers candidats. En effet, l'algorithme calcule

la différence en valeur absolue entre la première probabilité pignistique et la deuxième probabilité pignistique et la compare au seuil d'indécision fixé à 0.15. Ici, la différence étant inférieure au seuil d'indécision, l'algorithme décide le résultat *indécis*. En proposant de calculer le rapport entre le premier et le deuxième maximum des probabilités pignistiques, le problème lié au seuil d'indécision est résolu car la valeur issue du rapport est plus habilité à rester inférieure au seuil d'indécision que la valeur issue de la différence.

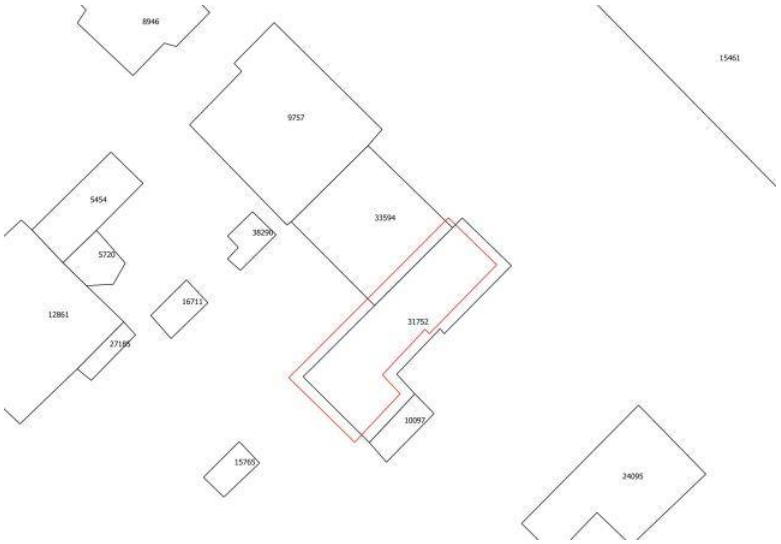


FIGURE 5.9 – appariement incorrect (*indécis* selon l'algorithme) : ProbPignisticMax= 0.1082 (qui correspond à la valeur "max") et ProbPignisticSecond= 0.00014, différence= 0.10806 < seuilIndecision(dans mon cas seuilIndecision=0.15), avec les valeurs suivantes sur les critères d'appariement : DR= 0,391m ; DA = 0,27 rad ; DS= 0,314 ; DH= 1,0074m

5.2.2 Résultats de l'appariement de données surfaciques

Nous rappelons que les jeux de données sur lesquelles portent notre étude ont été extraits sur une zone d'étude dans le département 94 (en Ile-de-France) sur le bâti de la BDTOPO d'une part et les bâtiments d'OSM d'autre part. Dans un processus d'appariement, pour chaque objet géographique A, appartenant à BDTOPO, l'algorithme d'appariement multi-critères recherche les candidats afin de choisir le meilleur candidat pour chaque objet géographique selon l'approche détaillée dans la sous-section 5.2.1.2. Le seuil de sélection des candidats est fixé empiriquement à 30 m. À la fin de l'appariement des données, chaque objet de BDTOPO est classé dans l'une des trois catégories suivantes : *apparié*, *non-apparié*, et *indécis*. La figure suivante 5.10 illustre les trois catégories de résultats possibles pour un appariement multi-critères basé sur la théorie de croyance :

- Appariés (matched en anglais) : l'objet **A** est apparié avec l'objet **X** d'OSM (en marron) car en terme de forme et de position, les deux objets se ressemblent et se superposent.
- Indécis : L'objet A de Bdtopo se situe en terme de position, entre deux objets OSM bien qu'en terme de forme l'un des deux objets géographiques d'OSM ressemble

d'avantage à l'objet A Bdtopo. L'algorithme a eu du mal à départager les deux objets géographiques.

- Non-Appariés : Après confrontation des deux jeux de données, on observe que l'objet A Bdtopo ne dispose pas d'objet homologue car il n'y a aucun objet géographique parmi les candidats, qui ressemble à l'objet A Bdtopo et non plus qui se superpose.

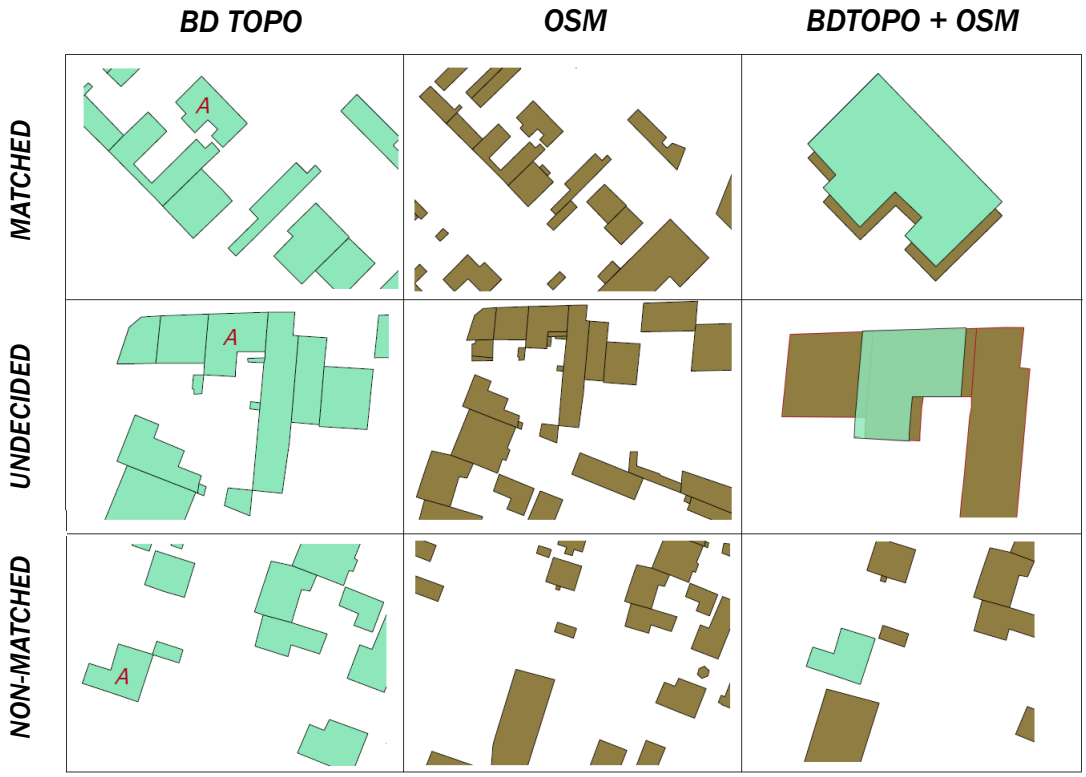


FIGURE 5.10 – les trois cas de résultat de l'algorithme d'appariement : entité A de BDTopo (à gauche), les candidats d'OSM (au centre), le résultat de l'appariement (à droite), respectivement pour le cas *apparié* (en haut), *indécis* (au milieu) et *non-apparié* (à droite).

Les résultats d'appariement obtenus pour le jeu utilisé dans la zone d'étude sont résumés dans le tableau 5.2.2selon les trois cas :
Parmi les 29152 objets du jeux de données de la BDtopo, 22 989 ont été appariés, 1143

type of matching	*Number*
matched	22989
non-matched	1143
undecided	5020

TABLE 5.5 – les résultats d'appariement du jeu de données surfaciques de la BDTopo

objets n'ont pas été appariés. Enfin, pour 5020 objets, l'algorithme n'a pas pu décider. On précise que tous les objets à apparier ont eu des candidats d'OSM. Pour le cas des indécis, la décision d'indécision apparait lorsque le premier max et le deuxième max des

probabilités pignistiques sont proche (écart entre les deux premières hypothèses, y compris l'hypothèse non-appariés est faible).

5.2.3 Validation de l'appariement

A l'issu de l'appariement, nous n'avons pas un lien d'appariement, mais une suite d'enregistrements contenant entre autre, les identifiants d'un objet BDTOPO et d'un objet candidat OSM, et une décision relative à leur appariement (*true*, *false* et *indécis*). Un seul parmi ces enregistrements porte la décision *true* et concrétise l'appariement qu'il y a eu entre les deux objets à apparier. A partir du fichier d'enregistrements des résultats d'appariement, nous visualisons un échantillon de paires d'objets appariés. Pour cela, on utilise un plugin [3] développé dans le laboratoire LASTIG qui affiche un groupe d'objets OSM autour d'un objet BDTOPO. De cette façon, nous pouvons vérifier visuellement les liens d'appariement. Voici un exemple affiché dans le plugin avec comme résultat :

- Appariement correct : l'objet OSM retrouvé et apparié par l'algorithme correspond bien à l'objet BDTOPO. Cet appariement s'avère correct après vérification comme illustré dans la figure 5.11.

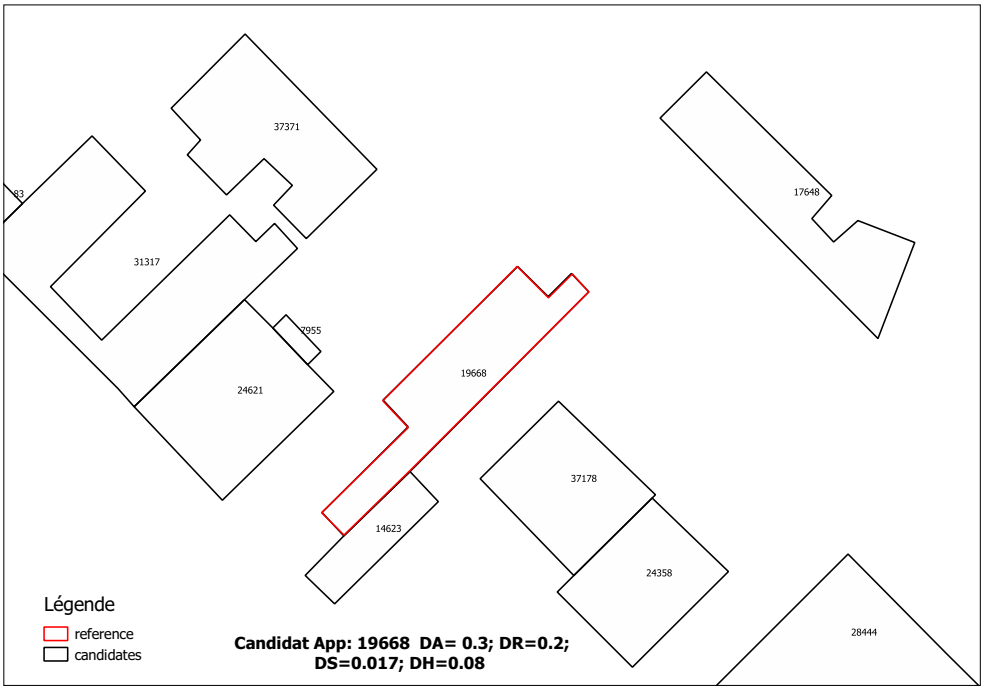


FIGURE 5.11 – Appariement correct : L'objet BDTOPO s'apparie avec l'objet OSM ID=16668 avec les valeurs suivantes sur les critères d'appariement : DA= 0.3 rad ; DR=0.2 m ; DS=0.017 ; DH=0.08 m

Pour que la vérification soit suffisante pour valider les résultats d'appariement, on doit s'assurer que notre échantillon soit représentatif de tous les résultats. Pour cela, nous nous

3. https ://github.com/mdvandamme/VisuValideMultiCriteriaMatching/

sommes intéressés à la question de la taille de l'échantillon sur lequel nous souhaitons calculer la proportion $\hat{p}$ *d'objets correctement appariés* ((objets correctement appariés) / (tous les objets appariés)). Nous savons que la valeur réelle de $\hat{p}$ se trouve dans l'intervalle $\hat{p} \pm \varepsilon(p, \alpha, N)$ avec la *marge d'erreur* donnée par l'équation suivante 5.1 :

$$\varepsilon(p, \alpha, N) = z_{\alpha/2}\sqrt{\frac{p(1-p)}{N}} \tag{5.1}$$

où N est le nombre de données et p est la valeur estimée de la proportion relative au N données, α est le risque (complémentaire du niveau de confiance), $z_{\alpha/2}$ est le nombre d'écarts types associé au risque α, en considérant que l'estimateur suit une loi normale.

Nous avons procédé à la vérification manuelle sur un premier échantillon. Puis nous avons calculé la proportion des objets correctement appariés. On obtient une valeur approximative de l'estimateur égale à $0,9759$ avec une marge d'erreur a priori inférieure ou égale à 0.1%.

Par la suite, nous essayons de calculer la taille nécessaire d'un échantillon représentatif qui réduirait la valeur de la marge d'erreur avec un niveau de confiance de 95% (risk $\alpha = 5\%$). En résolvant l'équation 5.1, nous obtenons une valeur $N = 174$ (avec $z_{\alpha/2} = 1.96$ et $p = 0.9759\text{-}0.1\text{=}0.8759$, $\varepsilon = 0.1$). A ce stade, nous reprenons la vérification manuelle en rajoutant à l'échantillon, 74 paires d'objets appariés. La vérification finale renvoie une valeur d'estimateur égale à $\hat{p} = 0,9885$.

Pour fournir une validation robuste de l'estimateur, la valeur de la marge d'erreur est recalculée à l'aide de l'équation 5.1 avec la nouvelle valeur de $\hat{p} = 0,9885$ et de $N = 174$. Ainsi, nous obtenons une estimation plus précise de la *proportion d'objets correctement appariées* égale à 0,9885 avec une marge d'erreur égale à 0.15% avec un niveau de confiance de 95%. Nous pouvons donc dire que la *Précision* de l'algorithme d'appariement est égale à $98,85$ $(+/-1,5)$.

Chapitre 6

Inférence de la qualité extrinsèque à partir des données intrinsèques

Sommaire

L'objectif de ce chapitre est d'expérimenter des méthodes statistiques permettant de mettre en relation des mesures obtenues à partir d'indicateurs mesurant la qualité extrinsèque, issues de l'appariement des données, avec celles d'indicateurs intrinsèques calculés uniquement sur les objets OSM appariés. La relation s'étudie à travers des méthodes d'apprentissage effectuant une régression (pour estimer les valeurs des indicateurs extrinsèques) ou une classification (pour séparer les objets en plusieurs classes de qualité). En premier lieu, nous étudions l'ensemble des méthodes définissant un modèle de régression à l'échelle des bâtiments individuels. Puis, nous étudions des méthodes prenant en compte des indicateurs à l'échelle mésoscopique, en considérant des agrégations de bâtiments, afin de mettre davantage en avant les caractéristiques propres au voisinage d'un bâtiment. En second lieu, dans une logique de relaxation du problème d'apprentissage, nous remplaçons le problème de régression par un problème de classification dans le but de détecter une qualité du bâti plutôt que d'évaluer numériquement la qualité extrinsèque.

6.1 Inférence de la qualité extrinsèque à partir des données intrinsèques avec des indicateurs locaux

6.1.1 Méthode d'inférence de la qualité extrinsèque à partir des données intrinsèques à l'échelle du batiment

6.1.1.1 Élaboration des indicateurs intrinsèques

Pour pouvoir inférer la qualité extrinsèque, nous proposons une démarche de définition des indicateurs intrinsèques. Ces indicateurs constituent la base de notre proposition de l'évaluation de la qualité en l'absence de données de référence. Par exemple, en s'inspirant entre autres, des travaux de Girres (2012a), notre réflexion a porté donc sur l'appréhension de ce qui pourrait nous révéler des traces d'une mauvaise saisie de données affectant la forme et la position des bâtiments. A la suite de notre analyse sur les origines des imperfections susceptibles de se produire à la saisie d'un bâtiment, nous dressons une liste d'indicateurs intrinsèques pertinents. Nous expliquons pour chaque indicateur en quoi il peut indiquer la mauvaise saisie d'un bâtiment.

Ainsi, nous distinguons deux types d'indicateurs : ceux que nous classons dans une catégorie nommée **hypothèse** et ceux classés dans une catégorie nommée **expérience**. On entend par **hypothèse** l'ensemble des indicateurs formulés sous l'angle d'une réflexion rationnelle. Nous avons effectivement voulu exprimer à travers ces indicateurs, une sorte de marqueur d'un élément qui est susceptible selon un raisonnement bien précis, d'expliquer une mauvaise saisie d'un bâtiment. En confrontant ces indicateurs aux données d'entrainement et du test, nous souhaitons parvenir à mieux prédire les nouvelles valeurs d'une variable d'étude (une mesure extrinsèque). Quand le modèle ne retient pas certains indicateurs d'hypothèses, il est intéressant de comprendre en quoi nos hypothèses n'étaient pas justes, malgré l'hypothèse de la pertinence **a priori** de ces indicateurs. En guise d'indicateurs d'hypothèses, on cite les indicateurs : **outlier (out)**, **rapport-long-min-long-max (lme)**, **longueur-minimale (lmn)**, **longueur-maximale (lmx)** et **q-reconstruct (qrc)**.

- **outlier (out)** : cet indicateur mesure le degré d'éloignement d'un sommet par rapport aux autres et à quel point il pourrait être considéré comme une valeur aberrante si on considérait l'ensemble des sommets du polygone comme une distribution statistique. Pour chaque sommet, on calcule une distance moyenne qui le sépare des autres sommets. Le sommet ayant la plus grande distance moyenne est sélectionné. Cette valeur est divisée par la longueur moyenne des segments du polygone. Plus la valeur est grande par rapport à 1, plus elle indique que le sommet en question pourrait être un point aberrant, et le résultat d'une erreur de saisie.

- **rapport-long-min-long-max (lme)** : cet indicateur est défini comme le rapport de la longueur la plus courte sur la longueur la plus grande des côtés d'un bâtiment. Ce rapport est d'autant plus proche de 0 qu'il exprime une grande irrégularité de forme qui peut suspecter une mauvaise qualité de saisie.

- **longueur-minimale (lmn)** : cet indicateur mesure combien de fois le segment le plus court du polygone est plus court que la longueur moyenne des segments du polygone. Il se calcule comme étant le rapport de la longueur du segment le plus court du polygone sur la longueur moyenne des segments de la polygone. Il prend des valeurs

inférieures ou égales à 1. Un faible rapport peut indiquer une erreur de saisie sur le segment en question.

- **longueur-maximale (lmx)** : similairement à l'indicateur **longueur-minimale**, cet indicateur mesure combien de fois la longueur du segment le plus long du polygone est plus grand que la longueur moyenne des segments du polygone. Une grande valeur peut indiquer une erreur de saisie sur le segment en question.

- **q-reconstruct (qrc)** : cet indicateur mesure la qualité de la reconstruction du polygone en utilisant seulement certains sommets parmi les sommets du polygone. Pour un seuil de reconstruction de 80% de la forme du polygone, nous calculons la proportion du nombre de sommets nécessaires à la reconstruction de la forme du polygone par rapport au nombre total de sommets du polygone. Une valeur proche de 1 indique qu'il n'y a pas de redondance dans la saisie, et une valeur plus petite indique une saisie avec une plus petite granularité. La redondance dans la saisie (valeur proche de 0) exprime une mauvaise qualité de saisie.

Par ailleurs, par la catégorie **expérience**, nous évoquons tous les indicateurs qui ont déjà été testés dans des études antérieures ayant pour but de qualifier la forme des bâtiments, et qui d'après ces études apportent de l'information sur la définition de la qualité de saisie d'un bâtiment. Nous les ajoutons à nos descripteurs sans toutefois avoir une explication nette de leur intégration dans le modèle d'apprentissage. Nous espérons qu'ils se manifestent dans le processus de sélection des variables importantes. Pour les indicateurs sélectionnés, nous souhaitons formuler une hypothèse expliquant leur sélection et si possible les transformer en des indicateurs d'hypothèses. Au final, le modèle d'apprentissage doit avoir la capacité d'expliquer les valeurs existantes de la variable cible (la qualité extrinsèque) et de prédire la qualité de nouveaux objets, mais doit aussi avoir une valeur explicative dans la mesure du possible. Ainsi, le modèle permettra de formuler des connaissances à partir de l'information issue des données. Les indicateurs d'expérience sont : **compacité (cpc)**, **convexité (cvx)**, **élongation (elg)**, **angle droit (ragl)**, **orientation (ori)**, **périmètre (per)**, **area (are)**, **granularité (grn)** et **rectangulaire (rec)**.

- **compacité (cpc)** : cet indicateur mesure la compacité du polygone par rapport à l'aire d'un cercle ayant le même périmètre que celui du polygone en sachant que le cercle est la figure dont l'aire est maximale pour un périmètre donné en vertu du théorème isopérimétrique. La compacité s'obtient par le rapport entre la surface du polygone étudié et celle du cercle ayant le même périmètre que le polygone. Les valeurs vont de 0 à 1, de sorte qu'une valeur proche de 0 (faible compacité) reflète une forme allongée, et une valeur proche de 1 (grande compacité) reflète une forme compacte ou circulaire. Un rapport faible et une forme très allongée peut être le signe d'une erreur de saisie. Un rapport faible et une forme très allongée peuvent être le signe d'une erreur de saisie.

- **convexité (cvx)** : la convexité d'un polygone est une propriété dictant que l'on peut se déplacer dans un polygone selon une droite tout en restant dans le polygone. En grossier, les polygones convexes ont une forme arrondi plutôt qu'une forme creuse ou bosselée. Cet indicateur mesure à quel degré la forme du polygone ressemble à une forme pleine. C'est le rapport de la surface du polygone sur la surface de son

enveloppe convexe. Les valeurs vont de 0 à 1 (parfaitement convexe). Une forme creuse d'un batiment peut signaler une mauvaise saisie du batiment.

- **élongation (elg)** : cet indicateur mesure le degré de ressemblance de la forme du polygone avec un carré. Il s'agit du rapport de la largeur sur la longueur du plus petit rectangle englobant (PRE). Il tend vers 1 lorsque la forme se rapproche de celle d'un carré. Moins la forme du polygone occupe l'aire du rectangle englobant (valeur proche de 0), plus nous suspectons une mauvaise qualité sur la saisie du batiment.

- **angle droit (ragl)** : cet indicateur mesure le nombre d'angles approximativement droits et nous renseigne sur la régularité de la forme d'un polygone, qui peut indiquer une bonne représentation de la forme d'un polygone.

- **rectangulaire (rec)** : cet indicateur mesure la manière dont un polygone remplit son plus petit rectangle englobant (ou en anglais, Smallest Surounding Rectangle, SSR), et il est calculé comme le rapport de la surface du polygone à la surface du SSR. L'indicateur prend la valeur 1 lorsque l'aire du polygone est parfaitement égale à celle du SSR et tend vers **0** lorsqu'elle est trop petite par rapport à celle du SSR. Cela nous renseigne sur la forme rectangulaire du bâtiment. On peut s'attendre à ce qu'un sommet mal saisi puisse casser la régularité d'un bâtiment et créer une différence d'aire plus importante entre le polygone et son SSR.

- **orientation (ori)** : cet indicateur mesure l'orientation du plus long coté du PPR (ou en anglais SSR) du polygone par rapport à l'axe **Ox** s'exprimant sous la forme d'un angle en radian dans $[0; \pi]$. L'angle calculé est d'autant plus grand que le bâtiment est entaché d'une erreur de saisie.

- **périmètre (per)** : cet indicateur vaut la somme des longueurs des cotés du polygone. Une valeur trop grande de périmètre du batiment pourrait résulter de la saisie du regroupement de plusieurs bâtiments, et de ce fait, elle peut laisser suspecter une mauvaise qualité de saisie.

- **area (are)** : cet indicateur correspond à l'aire du polygone. Une valeur trop grande de surface du batiment pourrait résulter de la saisie du regroupement de plusieurs bâtiments, et de ce fait, elle peut être signe d'une mauvaise qualité de saisie.

- **granularité (grn)** : la granularité correspond à la densité linéaire des points sur le contour. Cet indicateur mesure la granularité d'un polygone. Il est calculé comme étant le quotient du nombre de sommets et du périmètre. Plus la valeur est grande, plus on pense que le bâtiment a été saisi avec précision.

Bien que notre expérimentation se limite qu'aux indicateurs définis ci-dessus, nous précisons tout de même que cette liste d'indicateurs demeure non exhaustive et des études ultérieures peuvent mettre en évidence la pertinence d'autres indicateurs. Pour ce qui suit, ces indicateurs intrinsèques sont considérés comme des variables explicatives du problème d'apprentissage (régression ou classification). Nous commençons par rechercher une régression à l'échelle *micro* (sur un bâtiment).

6.1.1.2 Élaboration d'un premier modèle de régression

Nous recherchons une méthode générique qui pourrait relier les indicateurs intrinsèques aux indicateurs extrinsèques. Ainsi, nous optons pour l'utilisation d'un modèle de régression linéaire multiple qui est un modèle de base de l'apprentissage automatique.

Tout d'abord nous formulons notre problème de régression. Nous utilisons comme variables explicatives du modèle, l'ensemble des indicateurs intrinsèques définis ci-dessus, tandis que les indicateurs extrinsèques dérivés de la mise en correspondance des données sont considérés comme des variables dépendantes du modèle. Pour chaque bâtiment apparié issu de la base OSM, nous voulons associer une liste de valeurs issues de l'implémentation des variables explicatives et une valeur relative à chacune des variables dépendantes à savoir la distance radiale, distance angulaire, distance de Hausdorff et distance surfacique.

Puis, nous établissons le modèle de régression qui définit une relation estimée entre chaque variable dépendante et l'ensemble des variables explicatives. Plus généralement, la régression permet de répondre aux problèmes suivants :

- Identifier les variables explicatives qui sont associées à la variable dépendante.
- Prévoir les valeurs inconnues de la variable dépendante.

Le modèle de régression obtenu permettra à son tour, de prédire ou estimer une nouvelle valeur de la variable dépendante en prenant en entrée les valeurs explicatives calculées par exemple sur un nouveau bâtiment. Il doit tout de même être validé statistiquement à travers une analyse de régression.

Mathématiquement parlant, nous souhaitons expliquer une variable quantitative Y (exemple : distance radiale) en fonction de p autres variables $X_1, ..., X_p$ (rec, lme, lmx, $...grn$) Ainsi,
◦ Y est la variable à expliquer,
◦ $X_1, ..., X_p$ sont des variables explicatives.

Du point de vue pratique, nous souhaitons ajuster un modèle pour expliquer Y en fonction de $X_1, ..., X_p$.

Les données à traiter constituent des observations des ces variables. Ce sont n observations de $(Y, X_1, ..., X_p)$ notées $(y_1, x_{1,1}, ..., x_{p,1}), ...(y_n, x_{1,n}, ..., x_{p,n})$. Les données se présentent généralement sous la forme d'un tableau :

Si nous envisageons une relation linéaire entre Y et $X_1, ..., X_p$, on peut considérer le modèle de régression linéaire multiple suivant : il existe $p+1$ coefficients inconnus $\beta_0, ..., \beta_p$ tels que :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p + \epsilon \tag{6.1}$$

où ϵ est une quantité représentant une somme d'erreurs.

Y	X_1	$\dots$	X_p
y_1	$x_{1,1}$	$\dots$	$x_{p,1}$
y_2	$x_{1,2}$	$\cdots$	$x_{p,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{1,n}$	$\dots$	$x_{p,n}$

FIGURE 6.1 – Illustration du format de données.

L'objectif est donc d'estimer les coefficients inconnus $\beta_0, ..., \beta_p$ à l'aide des données afin de prédire la valeur de Y pour une nouvelle valeur de $(X_1, ..., X_p)$.

Pour cela, on modélise les variables considérées comme des variables aléatoires réelles (*var*) définies dans un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$.

$\forall i \in 1, ..., n$,
∘ $(x_{1,i}, ..., x_{p,i})$ est une réalisation du vecteur aléatoire réel $(X_1, ..., X_p)$,
∘ sachant que $(X_1, ..., X_p) = (x_{1,i}, ..., x_{p,i})$, y_i est une réalisation de :

$$Y_i = \beta_0 + \beta_1 x_{1,i} + ... + \beta_p x_{p,i} + \epsilon_i \tag{6.2}$$

où ϵ_i est une *var* modélisant l'erreur de régression de Y_i.
On appelle erreurs les var $\epsilon_i, ..., \epsilon_n$

Nous remarquons pour tout $x = (x_1...., x_p) \in \mathbb{R}^{\prime}$, sous l'hypothèse que $\mathbb{E}(\epsilon|(X_1, ..., X_p) = x) = 0$, le modèle de régression linéaire multiple (*rlm*) peut s'écrire comme :

$$\mathbb{E}(Y|(X_1, ..., X_p) = x) = \beta_0 + \beta_1 x_1 + ... + \beta_1 x_p \tag{6.3}$$

Ainsi, sachant que $(X_1, ..., X_p) = x$, la valeur moyenne de Y est une combinaison linéaire de $(x_1, ..., x_p)$.

En écriture matricielle, le modèle de *rlm* s'écrit sous la forme :

$$Y = X\beta + \epsilon \tag{6.4}$$

où $X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,n} & \cdots & x_{p,n} \end{pmatrix}$, $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$, $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$.

L'estimateur $\hat{\beta}$ des moindres carrés ordinaires (EMCO) de β est :

$$\hat{\beta} = (X^t.X)^{-1} X^t Y \tag{6.5}$$

Il est construit de sorte que l'erreur d'estimation entre $X\hat{\beta}$ et Y (dans le cas où les erreurs ont les mêmes variances) soit la plus petite valeur possible au sens $||.||^2$:

$$\underset{\beta \in \mathbb{R}^p+1}{\text{Argmin}} ||Y - \beta X||^2 \tag{6.6}$$

où ||.|| désigne la norme euclidienne de $\mathbb{R}^p$.

6.1.1.3 Analyse et raffinement du modèle de régression

Notre démarche en analyse de régression a pour objectif d'étudier la significativité du modèle de régression ainsi établi. De manière générale, l'analyse de la régression peut servir à :

- Analyser la performance de la régression (par validation croisée par exemple), avec des indicateurs type RMSE.
- Comprendre la relation entre les variables dépendantes et explicatives en analysant le sens de la régression, avec une discussion sur les variables explicatives choisies, celles laissées de côté, le signe et la valeur des coefficients.

Pour ce faire, rappelons la formulation d'un modèle des moindres carrés ordinaires utilisant l'équation suivante pour une observation i :

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon_i \tag{6.7}$$

avec Y_i = valeur observée de la variable dépendante au point i ; β_0 = interception avec l'axe des y (valeur constante) ; β_n = coefficient de régression ou pente pour la variable explicative n au point i ; x_n = valeur de la variable n au point i ; ϵ_i = erreur de l'équation de régression pour l'observation i.

Certains des éléments donnés par le modèle de régression interviennent dans l'analyse de la régression. En effet, l'analyse des résultats d'un modèle de régression repose sur un certain nombre d'éléments en guise d'exploration de la régression. Parmi ces éléments, nous notons le coefficient de régression relatif à une variable explicative, qui indique l'influence de cette variable explicative sur la valeur d'une variable dépendante donnée. L'ensemble des valeurs de pente (coefficients de régression) peuvent être comparées pour déterminer l'influence relative de chaque variable explicative sur la variable dépendante. Plus la valeur de pente est grande en valeur absolue, plus son influence est grande.

Plus précisément, notre analyse de régression doit vérifier si la variable X_j est utile dans notre modèle. Pour cela, nous souhaitons tester une hypothèse nulle de la forme $H_0 : \beta_j = 0$ contre l'hypothèse alternative $H_1 : \beta_j \neq 0$. L'hypothèse nulle signifie qu'il n'existe aucune relation linéaire entre la variable dépendante et la variable explicative j. Ce qui implique qu'il est inutile de rajouter cette variable dans l'équation de la régression. En pratique, sous H_0, on calcule la valeur dite *t value* (la statistique de Student), avec $t\ value = \frac{\hat{\beta}_j}{\sigma_{\hat{\beta}_j}}$ qui suit une loi de Student à $n-p$ degrés de liberté, que l'on peut approcher par une loi normale quand n devient très grand. $\hat{\beta}_j$ est la valeur estimée du coefficient de régression, $\sigma_{\hat{\beta}_j}$ est l'écart-type estimé sur ce coefficient ; il doit être faible par rapport à $\hat{\beta}_j$ pour qu'on puisse privilégier l'hypothèse H_1. La valeur de la statistique de Student est d'autant plus grande que la variable est significative. Nous pouvons également déduire de la **t-value**, ce qui est communément appelé la **p-value**.

En effet, pour pouvoir définitivement rejeter l'hypothèse nulle et affirmer la significativité d'une variable explicative, nous observons la valeur de **p-value**. Par définition, sous la supposition que H_0 est vraie, la *p-value* correspond à la probabilité d'observer les données obtenues (ou au moins aussi extrêmes). Généralement, on rejette l'hypothèse nulle H_0 quand la *p-value* est inférieure à 0.05.

Enfin, pour analyser le modèle de régression, nous nous basons sur la valeur du coefficient de détermination ou la part de variance expliquée par le modèle. Ce coefficient de détermination est noté R^2 et a pour expression :

$$R^2 = 1 - \frac{\text{Var}(residus)}{\text{Var}(apriori)} \tag{6.8}$$

avec $residus = (y_i - \widehat{y_i})$ les erreurs non expliquées par le modèle de régression pour une observation i.

La valeur R^2 est un nombre compris entre 0 et 1, les valeurs les plus proches de 1 indiquant des modèles qui expliquent le mieux la variable cible. Une valeur R^2 égale à 1 désigne un modèle parfait, ce qui est hautement improbable dans des situations réelles, étant donné la complexité des interactions entre différents facteurs et la possibilité que le comportement des variables à expliquer dépendent aussi de variables non connues. On s'efforce par conséquent de créer un modèle de régression dont la valeur R^2 est la plus élevée possible, tout en acceptant que cette valeur ne soit pas proche de 1. Parfois, les modèles faisant intervenir un grand nombre de variables ont mécaniquement un R^2 plus élevé, même si certaines des variables en question sont peu significatives (ce qui limite l'intérêt du modèle). La valeur R^2 ajustée, qui est également comprise entre 0 et 1, tient compte des variables explicatives supplémentaires, ce qui atténue le rôle que joue le hasard dans le calcul. Ainsi, la valeur R^2 ajustée doit être utilisée pour les modèles qui utilisent de nombreuses variables explicatives ou pour comparer des modèles comportant différents nombres de variables explicatives.

À l'issue de l'analyse de la régression, nous parvenons à définir et retenir les variables significatives (ou utiles) qu'il faut inclure le modèle de régression, ayant une **p-value** assez faible. Nous arrivons également à déterminer la part de variance réduite par le modèle de régression en suivant l'ajustement de la régression. À présent, nous examinons la qualité du modèle de régression.

À travers cette étude de la qualité du modèle, nous souhaitons retrouver les variables les plus influentes, et construire un modèle combinant au mieux parcimonie et qualité de l'ajustement. En effet, pour représenter un processus générant des données réelles, quel que soit le processus utilisé, il existe une perte d'information. Parmi donc les modèles-candidats, nous souhaitons choisir le modèle qui minimise cette perte d'information avec un nombre de paramètres suffisamment bas.

Ainsi, pour atteindre un équilibre entre ajustement et parcimonie, nous faisant appel au critère dit AIC afin d'obtenir le modèle le plus efficace. Ce critère ajoute donc une pénalisation au problème de régression. Cette pénalisation permet que le critère de décision soit minimisé pour un nombre restreint de variables, plutôt que d'avoir un critère qui s'améliore à chaque variable ajoutée. Nous procédons donc, à chaque étape, à une élimination de la variable la moins significative jusqu'à ce que l'AIC atteigne son mini-

mum.

Toutefois, cette démarche de sélection des variables reposant sur l'AIC a quelques défauts. En effet, quand deux modèles ont deux valeurs d'AIC proches, il difficile de savoir lequel choisir. Une des suggestions dans ce cas de figure, consiste par exemple à collecter plus de données pour espérer que leur distinction soit plus nette. Or, une modification infime de données peut engendrer un nouveau sous-ensemble des variables. Ainsi, l'ordre de retrait des variables serait altéré ainsi que le sous-ensemble de variables finalement retenu.

Pour remédier à cela, nous estimons qu'il est plus exacte d'utiliser une méthode de régularisation plus robuste que l'AIC dans le cadre de la sélection des variables pertinentes et nécessaires pour notre modèle de régression. Nous adoptons ainsi la méthode de pénalisation LASSO. Par sa pénalisation qui peut contraindre des coefficients de régression à prendre des valeurs faibles ou s'annuler, et peut réduire le nombre de paramètres pour atténuer la variance de l'estimateur, cette méthode évite le sur-apprentissage. Elle permet au final de sélectionner uniquement les variables les plus pertinentes de sorte à obtenir un modèle ayant un minimum de variables explicatives et qui peut être facilement interprétable. La méthode LASSO permet de sélectionner les variables pertinentes par une analyse visuelle comme déjà expliqué dans 3.4.2.3.

Afin de garantir la meilleure qualité du modèle de régression obtenu par la méthode LASSO, nous adoptons une approche de **validation-croisée** qui consiste, d'une part à entraîner le modèle sur une partie de l'échantillon pour produire l'estimateur, puis à évaluer la performance du modèle obtenu avec l'autre partie de l'échantillon. La validation croisée détermine l'erreur de test ou l'erreur de prédiction qui est l'erreur quadratique moyenne (**MSE pour Mean Square Error**). Nous avons utilisé une validation croisée à *2 folds*. A partir du MSE, on peut calculer le R^2 :

$$R^2 = \frac{\sum_i (\widehat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\sum_i (y_i - \widehat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} = 1 - \frac{\text{MSE}}{\text{Var}(Y)} \tag{6.9}$$

Dans un soucis de disposer d'une valeur stable R^2, nous devons la situer dans un intervalle de confiance en passant par un calcul d'incertitude. Ainsi, nous faisons appel à la méthode de **bootstrap** statistique (Singh et Xie, 2008) qui consiste à créer de nombreux échantillons dont chacun est divisé en partie d'entrainement et de test. Dans une seconde expérimentation (afin de renforcer la validation croisée), nous répétons l'opération de **bootstrap** 1000 fois et calculons chaque fois une valeur de R^2. Nous analysons la distribution des valeurs de R^2 en choisissant la valeur inférieure de la borne de l'intervalle de confiance à 95%. La part de variance expliquée exprime la capacité du modèle de régression à estimer une valeur numérique de la qualité extrinsèque sur un bâtiment donné. En d'autres termes, il s'agit d'une estimation à $(1 - R^2) * 100\%$ près de l'erreur relative attendue sur la forme ou la position d'un bâtiment. La valeur estimée peut à son tour, être définie dans un intervalle de confiance dans lequel elle est censée fluctuer. Plus le taux de variance expliquée est élevé, plus la prédiction est précise avec un intervalle de confiance plus étroit.

Enfin, notre démarche s'achève par une étape de validation du modèle de régression retenu. Cela passe par la validation d'un certain nombre d'hypothèses connues sous l'appellation **hypothèses de régression linéaire multiple** à savoir **l'hypothèse de normalité**,

l'hypothèse d'homoscédasticité et **l'hypothèse de non-multicolinearité** ainsi que quelques tests statistiques. L'hypothèse de **non-multicolinearité** suppose que les différentes variables explicatives ne sont pas liées les unes aux autres tandis que **l'hypothèse d'homoscédasticité** vérifie que les erreurs ne sont pas corrélés. **L'hypothèse de normalité** exprime à quel point la variable suit une loi normale.

6.1.2 Résultats

6.1.2.1 Résultats d'estimation des indicateurs extrinsèques à partir des indicateurs intrinsèques à l'échelle du bâtiment

Dans un objectif de déduire les variables dépendantes (indicateurs extrinsèques) des variables explicatives (indicateurs intrinsèques), nous recherchons un modèle de régression qui met en évidence l'existence d'une relation significative entre un indicateur extrinsèque et un groupe d'indicateurs intrinsèques. La significativité de la corrélation d'une variable explicative est atteinte lorsque la valeur de la *p-value* est inférieure à 0,05%.

Tout d'abord, nous effectuons une régression linéaire standard multiple reliant chacune des variables dépendantes (distance radiale, distance angulaire, distance surfacique et distance de Hausdorff) avec les 14 indicateurs définis dans la la section 6.1.1.1 et qui seront considérés comme variables explicatives.

Ensuite, par un processus descendant sur le calcul de la valeur de l'AIC, nous procédons à l'élimination des variables explicatives non significatives jusqu'à ce que nous obtenions un modèle parcimonieux, relatif à la plus petite valeur de l'AIC.

Enfin, pour affiner le modèle de régression linéaire couplé au critère d'AIC, nous développons une autre forme de régression en utilisant la méthode LASSO, qui devrait fournir un modèle parcimonieux, mais avec beaucoup moins de variables explicatives, avec pratiquement la même part de variance expliquée, réduisant ainsi la fonction de régression à quelques variables explicatives seulement. En régression LASSO, nous avons calculé la distribution des valeurs du taux de variance expliquée pour 1000 versions bootstrap en utilisant la validation croisée avec 70% des données utilisées pour l'entrainement et 30% pour la validation, sur un échantillon de 19 519 objets issus de l'appariement.

Pour ce qui suit, nous détaillons pour chaque indicateur, les résultats obtenus tout d'abord sur le modèle de régression linéaire standard, suivi des résultats du modèle appliqué au critère d'AIC, puis celles issues du modèle de régression avec la régularisation LASSO. A partir de la régression LASSO, nous identifions et énonçons les variables explicatives les plus importantes qui pourraient constituer le modèle de régression ultime avec une proportion de variance expliquée proche de celle du modèle de régression standard initial. Enfin, nous formulons l'équation de la fonction de prédiction de la qualité extrinsèque avec les coefficients des variables explicatives retenues.

Nous commençons par illustrer les résultats obtenus sur la distance angulaire choisie comme un indicateur de forme pour le bâti.

- **Distance angulaire** :
  Le tableau 6.1 représente les résultats de la régression linéaire standard pour l'indi-

cateur mesurant la distance angulaire (variable dépendante). Sur ce tableau, nous observons que presque toutes les variables sont significatives avec une *p-value* d'environ 10^{-16} témoignant largement de leur importance et de leur utilité dans le modèle de régression. Avec ce modèle de régression linéaire standard, nous estimons que la part de variance expliquée est égale à 29,33% par rapport à la variance totale sur la distance angulaire. Ce qui atteste que l'on peut prédire une valeur de distance angulaire, avec une incertitude moindre de 29,33% par rapport au cas où on se contenterait de prendre la valeur moyenne de cette distance sur le jeu de données.

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$1.11.10^{-1}$	$2.32.10^{-2}$	$1.74.10^{-6}$
a_{rec}	$-7.07.10^{-1}$	$2.51.10^{-2}$	$4.97.10^{-4}$
a_{lmx}	$-1.84.10^{-3}$	$2.78.10^{-3}$	0.507105
a_{lmn}	$3.87.10^{-2}$	$5.65.10^{-3}$	$***$
a_{lme}	$-1.053.10^{-1}$	$1.08.10^{-2}$	$***$
a_{out}	$4.44.10^{-2}$	$1.97.10^{-3}$	$***$
a_{cpc}	$-2.13.10^{-1}$	$2.09.10^{-2}$	$***$
a_{cvx}	$1.05.10^{-1}$	$4.37.10^{-2}$	$1.55.10^{-2}$
a_{elg}	$1.27.10^{-1}$	$8.38.10^{-3}$	$***$
a_{qrc}	$2.54.10^{-2}$	$7.62.10^{-3}$	$8.55.10^{-4}$
a_{ragl}	$-1.29.10^{-2}$	$1.01.10^{-2}$	0.201615
a_{ori}	$3.02.10^{-3}$	$1.19.10^{-3}$	$1.14.10^{-2}$

TABLE 6.1 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant |t|]$) pour la distance angulaire. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

Pour rendre parcimonieux le modèle de régression tout en gardant le maximum d'information utile, nous raffinons le modèle de régression avec un processus de sélection de variables significatives avec l'aide du critère d 'AIC. Sur la figure 6.2, nous observons que les variables explicatives moins significatives x_{ragl} (angle-droit) et x_{lmx} (segment-long) sont éliminées du modèle de régression ainsi que les variable x_{cvx} (convexité) et x_{rec} faiblement significatives.

Pour un intérêt d'interprétabilité, nous souhaitons réduire davantage le nombre des variables explicatives pour atteindre le nombre minimum conservant l'utilité du modèle. Nous faisons appel à la régularisation LASSO afin de contraindre certains coefficients à s'éliminer. La figure 6.2 exprime l'effet de la régularisation LASSO sur le modèle de régression linéaire standard pour la distance angulaire. Sur le graphique de droite de la figure, nous observons l'apparition des variables explicatives par ordre décroissant d'importance en fonction de la pénalité de la norme L_1. Sur le graphique de gauche de la figure, l'erreur quadratique moyenne de prédiction (erreur en validation croisée) est illustrée en fonction des valeurs choisies pour λ. En effet, plus la régularisation (coût sur les coefficients : la valeur de λ) est importante, plus la valeur de la norme L_1 est faible mais l'erreur quadratique moyenne sur la prédiction risque de s'accroitre. Tant que la valeur de la norme $L1$ est égale à 0, le modèle demeure vide et à mesure qu'on s'autorise à augmenter la valeur de la norme L_1, on assiste à l'apparition progressive des variables explicatives dans le modèle. On s'intéresse à

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$1.1439.10^{-1}$	$9.72.10^{-3}$	$***$
a_{lmn}	$3.88.10^{-2}$	$4.99.10^{-3}$	$7.74.10^{-15}$
a_{lme}	$-1.052.10^{-1}$	$8.69.10^{-3}$	$***$
a_{out}	$4.34.10^{-2}$	$1.67.10^{-3}$	$***$
a_{cpc}	$-2.08.10^{-1}$	$1.16.10^{-2}$	$***$
a_{elg}	$1.27.10^{-1}$	$6.20.10^{-3}$	$***$
a_{qrc}	$2.45.10^{-2}$	$7.54.10^{-3}$	$1.15.10^{-3}$
a_{ori}	$2.91.10^{-3}$	$1.19.10^{-3}$	$1.46.10^{-2}$

TABLE 6.2 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant |t|]$) pour la distance angulaire du modèle de régression linéaire standard après l'application du critère d'AIC. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

la valeur de λ (que nous noterons λ_{min} avec laquelle nous obtenons la plus petite erreur en validation croisée :

$$\lambda_{min} = \underset{\lambda \in \mathbb{R}^+}{\operatorname{argmin}} \, \mathbb{E}\Big[\big(Y - f_\lambda(X)\big)^2\Big]$$

où f_λ est le modèle de régression obtenue avec la pénalisation λ et où $f_\lambda(X)$ désigne sa prédiction effectuée sur un vecteur de variables explicatives X.

En pratique, le minimum de la norme L_2 (erreur en validation croisée) n'est pas atteint en un point bien défini λ_{min} ; du fait de l'incertitude entachant la détermination de l'erreur de prédiction par la validation croisée, la quantité λ_{min} est également *instable*. C'est pourquoi nous l'accompagnons d'un écart-type σ_λ. En vue de rendre potentiellement l'estimateur encore plus parcimonieux, le choix optimal de λ peut être déporté de λ_{min} vers : $\lambda^* = \lambda_{min} + \sigma_\lambda$, sans grande conséquence sur la capacité prédictive du modèle (erreur L_2). En contre-partie, la pénalisation sur la norme L_1 est renforcée, ce qui augmente les chances de retirer plus de variables explicatives du modèle.

Sur le graphique à gauche de la figure 6.2, λ_{min} est matérialisée par la droite en pointillés rouge, et λ^* est donnée par la droite en pointillés en bleu.

Ainsi avec une valeur de λ^* choisi par l'algorithme, le processus de régularisation LASSO nous donne la liste restreinte des variables explicatives. Sur le tableau 6.3, nous affichons uniquement les variables dont les coefficients n'ont pas été réduit à zéro.

À la lecture de la figure 6.2 mais aussi du tableau des coefficients 6.3, on peut observer les 5 premières variables importantes (*cpx* **: compacité,** *elg* **: élongation,** *rect* **: rectangularité,** *out* **: outlier et** *lmn* **: longueur-minimale)**. Nous remarquons que le modèle de régression basé sur la méthode LASSO reprend la variable *rect* (rectangularité) et *lmx* (longueur-maximale) dans son ensemble de variables significatives qui ont été éliminées auparavant par le critère d'AIC. Ce qui souligne la non-unicité

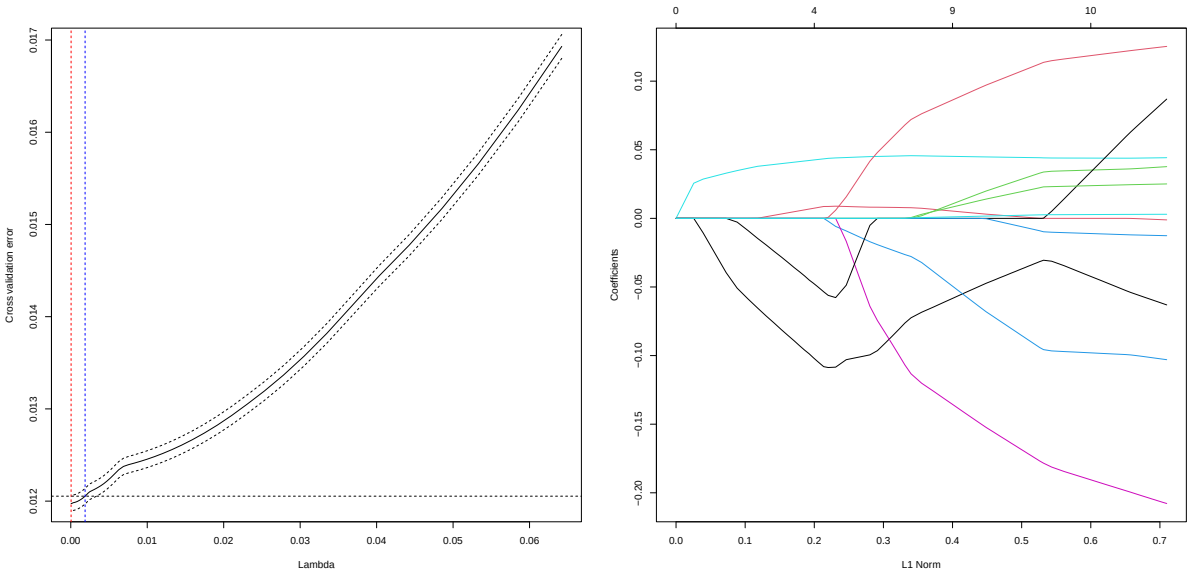


FIGURE 6.2 – modèle de régression LASSO pour la distance angulaire.

Coeffs. a	$\hat{a}$
a_{rec}	$-5.77.10^{-2}$
a_{lmx}	$5.18.10^{-3}$
a_{lmn}	$1.10.10^{-2}$
a_{lme}	$-5.04.10^{-3}$
a_{out}	$4.51.10^{-2}$
a_{cpc}	$-1.36.10^{-1}$
a_{elg}	$8.68.10^{-2}$
a_{qrc}	$8.60.10^{-3}$
a_{ori}	$1.11.10^{-3}$

TABLE 6.3 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance angulaire du modèle de régression LASSO.

de la configuration du modèle de régression proposée par un critère d'AIC. En effet l'ordre de retrait des variables peu ou non significatives peut engendrer l'obtention de plusieurs configurations du modèle de régression (en terme de variables explicatives contenues dans le modèle). C'est pourquoi nous avons basé notre sélection finale des variables explicatives sur celle issue des résultats de la régression LASSO. Aussi, la régularisation LASSO réduit t-elle les coefficients des variables *qrc* (qualité-reconstruction) et *ori* (orientation).

Ainsi nous formulons l'équation du modèle final de régression par une fonction de prédiction de la qualité extrinsèque Y (ici la distance angulaire) comme suit :

$$\widehat{Y} = 0.143 - 0.1364x_{cpc} + 0.0868x_{elg} - 0.0577x_{rec} + 1.10x_{lmn} + 0.0451x_{out} \quad (6.10)$$

Sur ce modèle de régression LASSO, nous avons obtenu une proportion de variance

expliquée à **27.34%** avec une borne inférieure de l'intervalle de confiance *L95* à **25.43%**. Nous pouvons donc affirmer que la proportion de variance expliquée par la régression LASSO est supérieure à **25.43%** avec un risque d'erreur de 5%.

À l'issue de la détermination de l'expression de l'équation, nous sommes en mesure de prédire ou estimer une nouvelle valeur sur la variable dépendante qui exprime un écart de forme (une composante de la précision spatiale) d'un bâtiment sans devoir passer par une comparaison avec une base de référence. Il suffira uniquement de calculer des valeurs d'indicateurs intrinsèques sur le bâtiment en question. Ainsi, pour un bâtiment ayant les valeurs suivantes sur les **5** variables explicatives (dans le cas de la distance angulaire) : $X = (x_{rec} = 5.3 * 10^{-3}, x_{out} = 2.032, x_{elg} = 0.665, x_{cpc} = 0.8807, x_{lmn} = 0.229)$, le modèle donne la valeur prédite égale à 0.4005, variant dans l'intervalle de confiance à **95%** $[0.015; 0.44]$.

- **Distance radiale**
  Le deuxième indicateur de forme pour lequel nous avons cherché une relation linéaire avec les indicateurs intrinsèques est la distance linéaire. En régression linéaire standard comme illustré sur la figure 6.4, nous constatons que plusieurs variables explicatives ne sont pas significatives. Seules les variables *rec* (rectangularité), *lmx* (longueur-maximale), *lmn* (longueur-minimale), *lme* (rapport-segment-court-segment-long), *out* (outlier), *cpc* (compacité) et *elg* (elongation) ont un apport significative pour la constitution du modèle de régression avec une proportion de variance expliquée s'élevant à **14.2%**.

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$-4.57.10^{-2}$	$6.21.10^{-2}$	$4.62.10^{-2}$
a_{rec}	$4.72.10^{-2}$	$6.72.10^{-2}$	$***$
a_{lmx}	$2.72.10^{-2}$	$7.43.10^{-3}$	$2.48.10^{-4}$
a_{lmn}	$-6.52.10^{-2}$	$1.50.10^{-2}$	$1.56.10^{-5}$
a_{lme}	$1.49.10^{-1}$	$2.88.10^{-2}$	$2.55.10^{-7}$
a_{out}	$7.63.10^{-2}$	$5.26.10^{-3}$	$***$
a_{cpc}	$-9.51.10^{-1}$	$5.59.10^{-2}$	$***$
a_{cvx}	$4.48.10^{-2}$	$1.16.10^{-1}$	$7.00.10^{-1}$
a_{elg}	$2.04.10^{-1}$	$2.23.10^{-2}$	$***$
a_{qrc}	$-7.5.10^{-4}$	$2.03.10^{-2}$	$9.70.10^{-1}$
a_{ragl}	$2.46.10^{-1}$	$3.05.10^{-2}$	$9.82.10^{-1}$
a_{ori}	$8.57.10^{-4}$	$3.19.10^{-3}$	$7.88.10^{-1}$

TABLE 6.4 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant \lvert t \rvert]$) pour le cas de la distance radiale en régression linéaire standard. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

Bien que beaucoup de variables explicatives se soient révélées moins significatives, nous tentons de raffiner le modèle de régression par une sélection des variables importantes via le critère d'AIC. Le résultat de la sélection se présente comme suit (voir tableau 6.5) :

Sur le tableau 6.5, nous constatons qu'effectivement les variables non significatives

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$-2.45.10^{-2}$	$3.08.10^{-2}$	$4.24.10^{-2}$
a_{rect}	$4.92.10^{-2}$	$4.13.10^{-3}$	$***$
a_{lmx}	$7.23.10^{-2}$	$7.43.10^{-3}$	$2.36.10^{-4}$
a_{out}	$7.61.10^{-2}$	$5.23.10^{-3}$	$***$
a_{cpc}	$-9.40.10^{-1}$	$4.89.10^{-2}$	$***$
a_{elg}	$2.01.10^{-1}$	$2.08.10^{-2}$	$***$
a_{lmn}	$-6.48.10^{-2}$	$1.50.10^{-2}$	$1.65.10^{-5}$
a_{lme}	$1.47.10^{-1}$	$2.87.10^{-2}$	$2.71.10^{-7}$

TABLE 6.5 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant \lvert t \rvert]$) pour la distance radiale du modèle de régression linéaire standard après l'application du critère d'AIC. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

ont été éliminées du modèle de régression sans dégrader la qualité d'ajustement.

À la fin traitement sur notre modèle de régression, la régularisation LASSO ne retient que les variables explicatives les plus importantes en mettant à zéro les coefficients du reste des variables explicatives comme illustré sur la figure 6.6. Avec un λ^* sélectionné par l'algorithme, nous observons le détachement des 5 premières variables importantes es par ordre de décroissant à garder dans le modèle de régression (voir figure 6.3).
Cela nous amène à formuler une fonction de prédiction sur une valeur de la qualité

Coeffs. a	$\hat{a}$
a_{cpc}	$-4.35.10^{-1}$
a_{rec}	$7.24.10^{-2}$
a_{out}	$7.20.10^{-2}$
a_{elg}	2.53^{-2}
a_{lme}	$2.12.10^{-3}$

TABLE 6.6 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance radiale du modèle de régression LASSO.

extrinsèque pour le cas de la distance radiale comme suit :

$$\widehat{Y} = -0.0245 - 0.435x_{cpx} + 0.0724x_{rect} + 0.0720x_{out} + 0.0253x_{lmx} + 0.00212x_{lme} \quad (6.11)$$

Ainsi nous observons que le modèle obtenu avec la distance radiale fut moins significative que celui obtenu avec la distance angulaire. Néanmoins, dans le papier que nous avons publié récemment (Menereaux et al., 2022), nous avons mené des expérimentations pour étudier la sensibilité de la distance radiale.

Nous illustrons pour ce qui suit brièvement, les résultats du modèle de régression tout d'abord pour la distance surfacique puis pour la distance de distance de Hausdorff.

- **Distance surfacique**
  La table 6.7 représente les résultats obtenus pour le modèle de régression linéaire

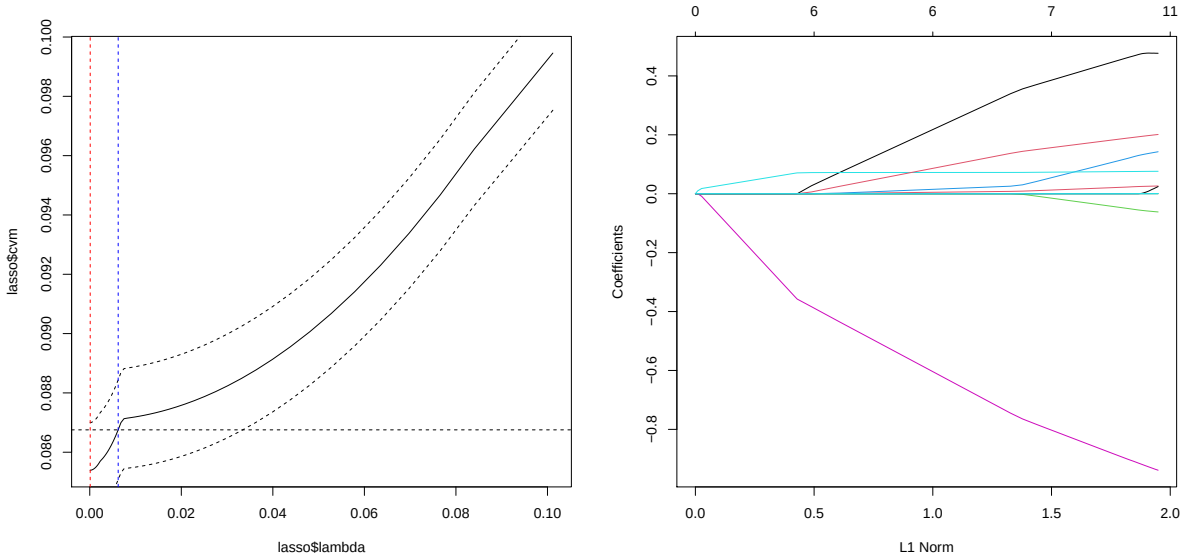


FIGURE 6.3 – modèle de régression LASSO pour la distance radiale.

standards obtenus pour la distance radiale.

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$6.77.10^{-2}$	$3.96.10^{-2}$	$8.72.10^{-2}$
a_{rec}	$-2.68.10^{-1}$	$4.28.10^{-2}$	$3.99.10^{-10}$
a_{lmx}	$-1.90.10^{-2}$	$4.73.10^{-3}$	$5.85.10^{-5}$
a_{lme}	$-2.42.10^{-3}$	$1.84.10^{-2}$	$8.95.10^{-1}$
a_{out}	$-1.04.10^{-2}$	$3.35.10^{-3}$	$1.88.10^{-3}$
a_{lmn}	$5.54.10^{-2}$	$9.61.10^{-3}$	$5.60.10^{-2}$
a_{cpc}	$-2.26.10^{-1}$	$3.56.10^{-2}$	$2.23.10^{-10}$
a_{cvx}	$5.94.10^{-1}$	$7.44.10^{-2}$	$1.59.10^{-15}$
a_{elg}	$3.96.10^{-3}$	$1.42.10^{-2}$	$7.81.10^{-1}$
a_{qrc}	$-1.26.10^{-2}$	$1.29.10^{-2}$	$3.30.10^{-1}$
a_{ragl}	$-1.28.10^{-1}$	$1.72.10^{-2}$	$1.03.10^{-13}$
a_{ori}	$-2.21.10^{-3}$	$2.03.10^{-3}$	$2.77.10^{-1}$

TABLE 6.7 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant \lvert t \rvert]$) pour le cas de la distance surfacique. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

En appliquant le modèle de régression initial au critère d'AIC, nous obtenons une sélection plus restreinte sur les variables explicatives (voir table 6.8).

Nous finissons par raffiner notre modèle de régression par une régularisation LASSO afin d'obtenir le modèle final de régression comme illustré sur la table 6.9. Nous retenons les 5 variables explicatives les plus importantes pour constituer notre modèle de régression à savoir **compacité,rectangularité, outlier longueur-maximale et longueur-**

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$6.63.10^{-2}$	$3.22.10^{-2}$	$3.99.10^{-1}$
a_{rect}	$-2.71.10^{-1}$	$4.21.10^{-2}$	$1.29.10^{-10}$
a_{lmx}	$-1.91.10^{-2}$	$3.83.10^{-3}$	$5.88.10^{-5}$
a_{out}	$-9.45.10^{-3}$	$3.18.10^{-3}$	$2.99.10^{-3}$
a_{cpc}	$2.13.10^{-1}$	$2.24.10^{-2}$	$***$
a_{cvx}	$5.78.10^{-1}$	$6.83.10^{-2}$	$***$
a_{lmn}	$-6.48.10^{-2}$	$1.50.10^{-2}$	$1.65.10^{-5}$
a_{agl}	$-1.13.10^{-1}$	$1.72.10^{-2}$	$3.49.10^{-14}$

TABLE 6.8 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant \lvert t \rvert]$) pour la distance surfacique du modèle de régression linéaire standard après l'application du critère d'AIC. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

minimale

Avec ce modèle final de régression, pour le cas de la distance surfacique, nous

Coeffs. a	$\hat{a}$
a_{cvx}	$1.917.10^{-1}$
a_{cpc}	$-1.2890.10^{-1}$
a_{agl}	$-1.1825.10^{-1}$
a_{rec}	$-6.801.10^{-2}$
a_{lmx}	-1.335^{-2}

TABLE 6.9 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance surfacique du modèle de régression LASSO.

parvenons à prédire une valeur sur la qualité extrinsèque seulement avec un taux de variance expliquée de **4%**. Ce qui témoigne d'une performance très faible sur la qualité de prédiction de ce modèle de régression basé sur la distance surfacique. l'équation suivante représente la fonction de prédiction du modèle de regression obtenu avec la distance surfacique :

$$\widehat{Y} = 1.162.10^{-4} + 0.191x_{cpx} - 0.129x_{cvx} - 0.118x_{agl} + -0.068x_{rec} - 0.0013x_{lmx} \quad (6.12)$$

- **Distance Haussdorf**

  Le modèle de régression basé sur la distance de Hausdorff est aussi très faiblement significatif comme son homologue obtenu avec la distance surfacique. Le taux en variance expliqué du modèle de régression est environ de **3%.** Nous donnons les résultats obtenus en différents processus de régression (régression linéaire standard, régression avec AIC, régression LASSO) de manière systématique.

  En régression LASSO, nous constatons qu'aucune variable explicative n'a été retenue. Ceci semble être normal étant donnée que le taux de variance expliquée qui était assez bas, la pénalisation minimale de LASSO a du rendre à zéro tous les coefficients

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$2.32.10^{-1}$	$2.18.10^{-2}$	$2.85.10^{-1}$
a_{rec}	$1.39.10^{-1}$	$2.35.10^{-1}$	$5.54.10^{-1}$
a_{lmx}	$8.65.10^{-3}$	$2.60.10^{-2}$	$7.40.10^{-1}$
a_{lme}	$-2.26.10^{-4}$	$1.01.10^{-1}$	$9.98.10^{-1}$
a_{lmn}	$5.85.10^{-3}$	$5.29.10^{-2}$	$9.11.10^{-1}$
a_{out}	$1.62.10^{-1}$	$1.84.10^{-2}$	$***$
a_{cpc}	$-3.45.10^{-1}$	$1.96.10^{-1}$	$7.86.10^{-2}$
a_{cvx}	$3.80.10^{-1}$	$4.09.10^{-2}$	$3.53.10^{-1}$
a_{elg}	$1.74.10^{-1}$	$7.85.10^{-2}$	$2.65.10^{-2}$
a_{qrc}	$5.89.10^{-2}$	$7.14.10^{-2}$	$4.08.10^{-1}$
a_{agl}	$-3.58.10^{-1}$	$9.51.10^{-2}$	$1.63.10^{-4}$
a_{ori}	$-2.58.10^{-2}$	$1.12.10^{-2}$	$2.13.10^{-2}$

TABLE 6.10 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant |t|]$)pour le cas de la distance de Hausdorff. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

Coeffs. a	$\hat{a}$	σ_a	$\mathbb{P}[\geqslant \lvert t \rvert]$
Intercept	$1.55.10^{-1}$	$1.84.10^{-1}$	$3.97.10^{-1}$
a_{out}	$1.67.10^{-1}$	$1.32.10^{-2}$	$***$
a_{cvx}	$5.51.10^{-1}$	$2.51.10^{-1}$	$2.81.10^{-1}$
a_{elg}	$1.76.10^{-1}$	$7.61.10^{-2}$	$2.03.10^{-2}$
a_{agl}	$-3.57.10^{-1}$	$9.49.10^{-2}$	$1.69.10^{-4}$
a_{ori}	$-2.59.10^{-2}$	$1.12.10^{-2}$	$2.06.10^{-2}$

TABLE 6.11 – Illustration des valeurs estimées des coefficients (a) avec leur écart-type σ_a et la p-value ($\mathbb{P}[\geqslant |t|]$) pour la distance de Hausdorff du modèle de régression linéaire standard après l'application du critère d'AIC. Le symbole $***$ signifie que la p-value est inférieure à 2.10^{-16}.

des variables explicatives.

La fonction de prédiction basé sur le modèle de régression avec la sélection des variables explicatives avec AIC s'écrit de cette manière :

$$\widehat{Y} = 0.155 + 0.167x_{out} + 0.551x_{cvx} + 0.176x_{elg} - 0.357x_{agl} - 0.0259x_{ori} \qquad (6.13)$$

À l'issue des résultats d'uue régression basée sur les indicateurs de position, nous avons remarqué que les scores de performance de prédiction de la qualité extrinsèque (en terme taux de variance expliquée) étaient très faibles.

En somme, nous présentons les variables explicatives les plus importantes qui devraient figurer dans le modèle finale de régression pour mener une démarche explicative du modèle de régression pour chacune des mesures extrinsèques de la qualité.

- distance angulaire : *compacité*et*elongation*, *rectangualire*, *rapport-long-min-long-max*, et *outlier*.

- distance radiale :*compacité*, *rectangulaire*, *outlier*, *longueur-maxiamle* et *rapport-long-min-long-max*.

- distance surfacique :*compacité*, *convexité*, *angle-droit*, *rectangualire* et *longueur-maximale*.

- distance de Hausdorff :*outlier*, *convexité*, *élongation*, *angle-droit* et *orientation*

À l'issue de la présentation des résultats de la régression sur les différentes distances, nous détaillons les expérimentations entreprises pour étudier la validité du modèle de régression basé sur la distance angulaire.

6.1.2.2 Validation du modèle de régression

A l'issue de l'élaboration du modèle de régression LASSO, nous souhaitons étudier la validité de la régression. Pour ce faire, nous testons que les hypothèses permettant d'utiliser un modèle de régression multiple sont vérifiées.

Parmi celles-ci, nous relevons l'hypothèse de normalité. L'hypothèse de normalité suppose que notre variable dépendante est normalement distribuée. Pour vérifier cette hypothèse, nous établissons le diagramme Q-Q normal. Le Q-Q normal, ou diagramme quantile-quantile, est un outil graphique qui nous aide à évaluer si un ensemble de données est susceptible de provenir d'une distribution théorique telle qu'une distribution normale. Un diagramme Q-Q est un nuage de points créé en traçant deux ensembles de quantiles l'un par rapport à l'autre, avec les quantiles observés sur l'axe des y et leurs quantiles normaux théoriques sur l'axe des x. Si les deux ensembles de quantiles proviennent de la même distribution, nous devrions voir les points former une ligne à peu près droite.

Dans la figure 6.4, nous répartissons les données par quantiles normalisés de la distribution des résidus de nos données (quantiles observés). Puis, nous comparons ces valeurs à celles aux quantiles théoriques issus d'une loi normale. Nous observons que la plupart des nuages des points tombent sur la ligne sauf aux deux extrémités, où elles forment des queues lourdes. Hormis les queues lourdes aux extrémités, nous affirmons que le modèle de régression semble avoir une distribution gaussienne. Ce qui confirme l'hypothèse de normalité.

Par ailleurs, une autre hypothèse que nous souhaitons vérifier est, l'hypothèse d'homoscédasticité. Pour ce faire, nous nous intéressons à la dispersion des résidus. L'homoscédasticité est observée lorsque la dispersion des résidus est homogène sur tout le spectre des valeurs des variables explicatives. C'est une propriété souhaitable car si les résidus correspondent à des incertitudes sur les valeurs estimées, il n'y a aucune raison pour que la dispersion des résidus change avec les valeurs du prédicteur (variable explicative), d'autant plus que les résidus sont les erreurs non expliquées par la régression, constituant une variable aléatoire. Pour vérifier l'hypothèse, on représente les résidus estimés $\widehat{e}_i = Y_i - \widehat{Y}_i$ (Residuals) en fonction des valeurs estimées notées $\widehat{Y}_i$ pour la valeur observée notée Y_i. La technique d'estimation utilisée suppose que les résidus estimés ont une variance constante (non dépendante de i).

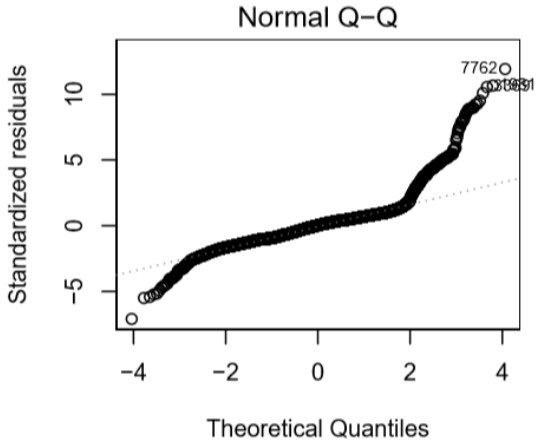


FIGURE 6.4 – Normal Q-Q pour le cas de la distance angulaire

Ainsi, dans la figure 6.5, où nous représentons les résidus estimés en fonction des valeurs ajustées, nous pouvons voir que les résidus se dispersent de manière aléatoire indépendamment des valeurs ajustées. Cela montre que la variance des résidus est homogène et constante, et on en déduit que l'hypothèse d'homoscédasticité est vérifiée.

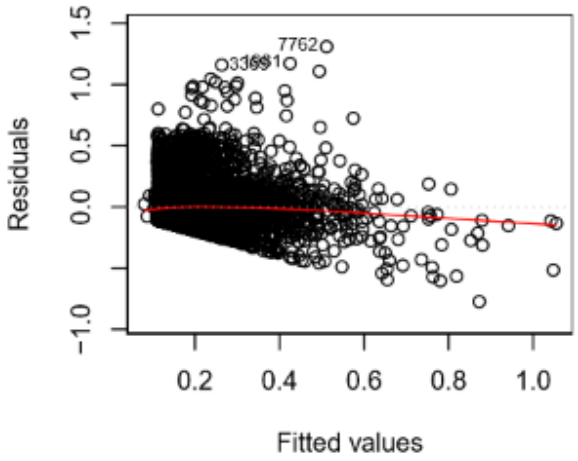


FIGURE 6.5 – Représentation graphique des résidus en fonction des valeurs ajustées

Pour renforcer l'hypothèse d'homoscédasticité, le test de Breusch-Pagan est utilisé pour déterminer la nature de la variance du terme d'erreur (résidus) : si la variance du résidu est constante quand la valeur de l'estimateur varie, alors nous avons l'homoscédasticité. Au contraire, si la variance varie, alors nous avons une hétéroscédasticité. En prenant l'hétéroscédasticité comme hypothèse nulle H_0, il suffit de vérifier si la *p-value* est inférieure à 5%, ce qui donne une *p-value* inférieure à 10^{-16}. Cela confirme l'hypothèse d'homoscédasticité (Zaman (2000)).

D'autre part, nous avons voulu voir si les résidus n'étaient pas auto-corrélés. Le test de Durbin-Watson est un test statistique conçu pour tester l'auto-corrélation des résidus

dans un modèle de régression linéaire. La technique d'estimation utilisée suppose que les résidus ne sont pas corrélés, de sorte que la statistique de Durbin-Watson (DW) doit être proche de la valeur 2. À la fin du test de Durbin-Watson, le DW= 2,0135 est calculé. Cela confirme la non-corrélation des résidus de notre modèle de régression et donc l'hypothèse d'indépendance des résidus Draper et Smith (1998).

En outre, sur la figure 6.5, nous observons visuellement que les résidus ne présentent aucune organisation particulière. Ce qui confirme l'hypothèse de linéarité.

La dernière hypothèse à vérifier est l'hypothèse de non-multi-colinéarité. Au sens strict, on parle de multi-colinéarité parfaite lorsque l'une des variables explicatives d'un modèle est une combinaison linéaire d'une ou plusieurs autres variables explicatives introduites dans le même modèle. L'absence de multi-colinéarité parfaite est l'une des conditions requises pour pouvoir valider un modèle linéaire. En termes non statistiques, il y a colinéarité lorsque deux ou plusieurs variables mesurent la *même chose*. L'approche la plus traditionnelle consiste à examiner le facteur d'inflation de la variance (**FIV**) Stine (1995). Les **FIV** estiment dans quelle mesure la variance d'un coefficient est **augmentée** en raison d'une relation linéaire avec d'autres prédicteurs. Par exemple, un **FIV** de 1.8 nous indique que la variance de ce coefficient particulier est supérieure de 80 % à la variance qui serait observée si ce facteur n'avait aucune corrélation avec d'autres prédicteurs.

Toutefois, il n'y a pas de consensus sur la valeur du FIV au-delà de laquelle il faut considérer qu'il y a multi-colinéarité. Certains auteurs suggèrent d'examiner plus en détail les variables dont le **FIV** est supérieure à 2,5. Dans notre cas, dans le modèle de régression étudié (cas de la distance angulaire), pour les variables explicatives $\mathbf{X} = (x_{grn}, x_{out}, x_{elg}, x_{cvx}, x_{lmn})$, les valeurs suivantes des **FIV** sont obtenues : $\mathbf{FIV} = (fiv_{grn} = \mathbf{1.046491}, fiv_{out} = \mathbf{1.507921}, fiv_{elg} = \mathbf{1.387913}, fiv_{cvx} = \mathbf{1.566782}, fiv_{lmn} = \mathbf{1.538345})$. Ces valeurs sont suffisamment faibles confirmant l'hypothèse de la non-multi-colinéarité.

À l'issue de la vérification des hypothèses par des tests statistiques, nous confirmons la validité de notre modèle de régression.

À travers cette étude de régression à l'échelle micro (du bâtiment), nous avons proposé une approche qui permet de dériver des indicateurs extrinsèques à partir d'indicateurs intrinsèques. En utilisant des résultats d'une comparaison de données robuste (appariement), la démarche a pu établir un modèle de régression estimé pour quatre indicateurs de la qualité extrinsèque en utilisant un panel de 12 indicateurs intrinsèques. Cela permet d'estimer une éventuelle précision géométrique et positionnelle relative d'un bâtiment. Bien que les résultats soient significatifs mais modestes, ils ont montré qu'il existe un signal permettant de détecter et de prédire la qualité extrinsèque. Pour pouvoir améliorer la qualité de prédiction, nous estimons qu'il convient de mener des recherches supplémentaires en tenant compte des caractéristiques propres au voisinage d'un bâtiment de sorte à redéfinir le modèle de régression sur une échelle d'agrégat des bâtiments à savoir à l'échelle **méso**.

6.2 Inférence de la qualité extrinsèque à partir des données intrinsèques tenant compte du voisinage

Une première façon d'améliorer la performance de la régression consiste à prendre en compte le contexte spatial en se basant sur une notion de voisinage pour laquelle les valeurs d'un indicateur extrinsèque peuvent être autocorrelées. Le voisinage peut être généré sur la base de considérations urbaines, telles que l'alignement avec la route, la régularité, la similarité et la proximité des bâtiments. Le voisinage ainsi schématisé donnera une structure spatiale. Ensuite, pour cette structure spatiale, les bâtiments qui s'y trouvent pourraient être agrégés et la valeur de l'indicateur extrinsèque peut être ré-estimée à cette nouvelle échelle plutôt qu'à l'échelle individuelle.

6.2.1 Méthode d'inférence de la qualité extrinsèque à partir des données intrinsèques à l'échelle d'un agrégat de bâtiments

L'objectif est de présenter la méthode que nous avons proposée et qui vise à définir un voisinage dans lequel nous montrons que les valeurs sur les indicateurs extrinsèques varient de la même manière (existence d'une auto-corrélation spatiale). Notre objectif consiste à proposer une méthode qui permet de détecter des structures homogènes sur lesquelles nous pourrons prédire une valeur d'indicateur extrinsèque valable pour un ensemble de bâtiments. Ces structures pourraient servir de base à une nouvelle régression prédisant la qualité des bâtiments à l'échelle de l'agrégat et non à l'échelle individuelle du bâtiment.
Pour ce faire, nous commençons par rechercher l'étendue d'une corrélation spatiale entre les observations d'une variable dépendante dans l'espace (étude d'un co-variogramme en fonction de la distance euclidienne). Dans la limite de distance établie dans laquelle il semblerait avoir une auto-corrélation spatiale, nous recherchons une manière de définir une structure spatiale et ses voisinages. Sur cette structure spatiale, nous évaluons l'existence d'une auto-corrélation spatiale. Enfin, nous agrégeons nos bâtiments au sein de la structure spatiale tant pour la variable dépendante que pour les variables explicatives pour évaluer la significativité d'un modèle de régression à l'échelle de la structure spatiale.

6.2.1.1 Étude de l'autocorrélation spatiale

Dans l'objectif d'améliorer la performance du modèle de régression réalisé à l'échelle du batiment, nous recherchons une échelle plus large que celle du batiment sur laquelle nous pouvons avoir une meilleure connaissance pour prédire un indicateur extrinsèque. Nous étudions d'abord une corrélation basée sur la distance euclidienne. Cela nous permet de définir la portée spatiale de l'auto-corrélation d'une variable donnée (les variables dépendantes). Puis dans la limite de la portée de l'auto-corrélation, nous proposons une manière de former une structure spatiale (éventuellement topologique) dans laquelle nous justifions l'existence d'une auto-corrélation spatiale pour cette variable.

Étudier l'auto-corrélation spatiale revient à mesurer la ressemblance des valeurs d'une variable dans une certaine distance. L'idée fondamentale, admettant que la nature n'est pas entièrement **imprévisible**, stipule que deux observations situées l'une près de l'autre devraient, en moyenne, se ressembler davantage que deux observations éloignées.

Prenons un exemple dans le domaine de la prospection minière où la géostatistique est souvent utilisée pour la problématique de l'estimation des gisements miniers. Soit trois points de localisation, x_0, x_1, et x_2 sur un site couvrant d'un réservoir de minerai, représentés dans la figure 6.6. On mesure une quantité de présence du minerai pour chacun de ces trois points de localisation.
La quantité au point x_1 devrait ressembler plus (en moyenne) à celle observée au point

FIGURE 6.6 – exemple illustratif de la variabilité d'un phénomène à différentes localisations

x_0 qu'à celle au point x_2 par principe de proximité. Notre but est donc d'estimer au-delà de quelle distance cette ressemblance n'est plus admise entre les valeurs de nos variables d'étude.

D'une manière générale, considérions un couple de points $(x_i, x_i + h)$, où x_i est une localisation quelconque de l'espace et $x_i + h$ est sa translatée d'un vecteur h. En général on suppose sans perte de généralité que le phénomène est isotrope. Ce qui conduit à assimiler h à une valeur scalaire (autrement dit $x_i + h$ désigne un point quelconque séparé de x_i par une distance h). En chaque point **x** du plan, nous mesurons une grandeur **Z(x)** assimilée à une variable aléatoire. La différence des quantités **Z** mesurées en 2 lieux **Z(x)-Z(x+h)** est une nouvelle v.a pour laquelle nous pouvons calculer la variance. Sous l'hypothèse de stationnarité du processus stochastique **Z**, cette variance nous renseigne sur la ressemblance des valeurs prises en des sites voisins d'une distance **h**. Intuitivement cette variance devrait être plus petite lorsque les localisations sont rapprochées (les valeurs se ressemblent plus en moyenne avec **h** petite) et plus grande lorsque les localisations sont éloignées (avec **h** plus grande). Ainsi étudier l'auto-corrélation des valeurs de **Z(x)** à différentes localisations de **x**, revient à étudier la variance de la v.a **Z(x) - Z(x+h)** pour plusieurs valeurs de **h**. Sous l'hypothèse où cette variance atteint un palier (potentiellement asymptotique) **C**, on définit le palier de l'autocorrélation comme la plus petite valeur de h permettant d'atteindre **95%** de **C**. On appelle variogramme, la demi-variance de la différence entre **Z(x) et Z(x+h)** :

$$\gamma(x, x+h) = \frac{1}{2}\mathrm{Var}(Z(x) - Z(x+h)) \tag{6.14}$$

L'outil constitué par le variogramme sert à extraire les 2 premières moments (moyenne et variance), l'espérance mathématique de la v.a Z(x) n dépend que de x (i.e. $\mathbb{E}[Z(x)] = m$), l'espérance des écarts vaut zéro (i.e. $\mathbb{E}[Z(x) - Z(x+h)] = 0$) et la covariance entre $Z(x)$ et $Z(x+h)$ ainsi le que variogramme $\gamma(h)$ ne dépendent pas de la localisation x mais seulement de h. Ainsi le variogramme et la covariance deviennent donc des fonctions dépendant uniquement de la distance séparant les points d'observation et non plus de leur localisation exacte.

Par ailleurs, cette hypothèse de stationnarité, nous permet une représentation plus aisée du variogramme. Autrement dit le variogramme en h se calcule par la moitié de la moyenne des carrées des écarts entre les valeurs prises en des sites distants de h.

$$\gamma(h) = \frac{1}{2}\text{Var}[Z(x) - Z(x+h)] = \frac{1}{2}\mathbb{E}[(Z(x) - Z(x+h))^2] \tag{6.15}$$

où x est le vecteur de coordonnées (1, 2 ou 3 coordonnées selon le cas), h est le vecteur distance.

Cette fonction γ, habituellement monotone[1] croissante en fonction de h, synthétise beaucoup d'informations concernant le comportement conjoint des variables aléatoires. Parmi celles-ci, on retient trois paramètres illustrés dans la figure 6.7 :

- **Portée a** : Distance où deux observations ne se ressemblent plus du tout en moyenne, elles ne sont plus liées (covariance nulle) linéairement. À cette distance, la valeur du variogramme correspond à la variance de la variable aléatoire **Z(x)**.
- **Palier** $\sigma^2 = C_0 + C$: Variance de la *v.a* (**Z(x)**). Ce sont les écarts les plus grands, en moyenne entre deux *v.a* , en l'occurrence ici **Z(x)** et **Z(x+h)**.
- **Effet de pépite :** C_0 : Variation à très courte échelle, erreurs de localisation, erreurs d'analyse et précision analytique. Lorsque $h = 0$ on a $\gamma(0) = \frac{1}{2}\text{Var}[Z(x) - Z(x)] = 0$ et non C_0. Par contre $\lim_{\epsilon \to 0^+} \gamma(\epsilon) = C_0$.

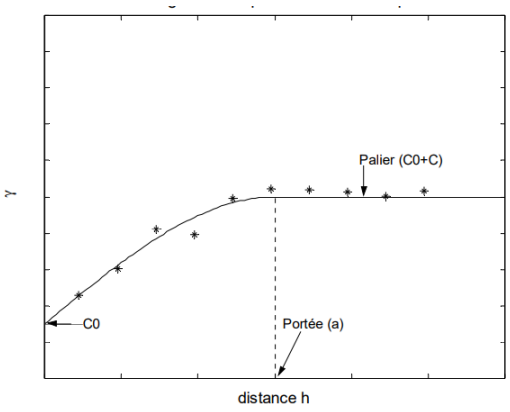


FIGURE 6.7 – variogramme experiemental et theorique. Source : extrait du *Cours GML6402 : Géostatistique, Chapitre 2 : le variogramme, Auteur : MARCOTE Denis, 2006* **https://cours.polymtl.ca/geo/marcotte/gml6402.html**

Lorsque le variogramme a un palier alors on peut établir le lien entre la valeur du variogramme pour la distance h et la covariance pour deux observations séparées de h :

$$\gamma(h) = \sigma^2 - C(h) \tag{6.16}$$

Nous remarquons que lorsque la portée est atteinte, il n'y a plus de covariance entre Z(h) et Z(x+h) (i.e. C(h) = 0 si $h \geq a$) quand le palier est atteint et C(h) ≈ 0 sinon, pour les modèles dit asymptomatiques, tandis que quand $h = 0$ $\gamma(0) = 0$ et donc $C(0) = \sigma^2$, variance du processus. Lorsqu'il y a un palier, les deux fonctions sont équivalentes en ce sens qu'elles fournissent la même information sur le processus.

1. un contre exemple notable est donné par le variogramme dit *à effet de trou*

En pratique, on définit une grille de maille régulière pour laquelle, on calcule la covariance pour un ensemble des valeurs de h espacées **d'un PAS de la maille**, d'une valeur située à la position x avec toutes les autres valeurs situées à $x + h$. Ce qui nous donne le co-variogramme expérimental.

Dans notre cas d'application, l'auto-corrélation spatiale est examinée tout d'abord à travers l'étude du co-variogramme des variables dépendantes. Pour cela, on trace la courbe du co-variogramme pour toutes les variables dépendantes (ou variables cibles) en fonction de la distance euclidienne du plan, pour des valeurs testées de **PAS** de 3 à 100 m, avec les résultats les plus probants à 10 m. Pour toutes ces variables, la portée de l'auto-corrélation semble exister significativement jusqu'à 800 m, puis décroit progressivement jusqu'à 2000 m. La figure 6.8 correspond au co-variogramme réalisé sur la distance angulaire :

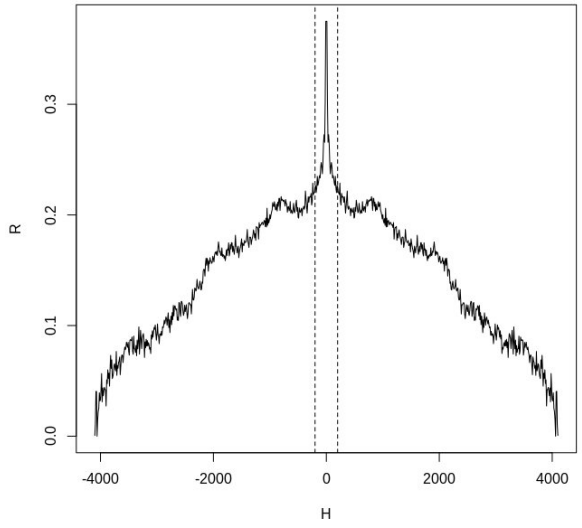


FIGURE 6.8 – Covariogramme sur la distance angulaire : on représente la covariance des valeurs de la distance angulaire (**R**) en fonction de la distance euclidienne (**H**).

Schématiquement, nous pouvons considérer que les processus stochastiques modélisant l'évolution des variables continues, dans notre cas, les distances angulaire, radiale, surfacique et Hausdorff sont composés majoritairement d'un bruit blanc et d'un bruit corrélé spatialement de variogramme gaussien (processus lisse). Pour les **4** variables cibles, le SNR (ratio de la variance du signal corrélé sur celle du bruit) est inférieur à l'unité. Les cas les plus favorables semblent survenir pour le cas de la distance angulaire DA (SNR 33%) et pour la distance radiale (SNR 50%).

Si nous souhaitons intégrer les variables cibles sur des emprises spatiales pour estimer la *qualité zonale*, on doit choisir un schéma d'agrégation des données. Au départ, nous ne disposons que de l'auto-corrélation en fonction de la distance euclidienne, ce qui peut permettre éventuellement de déterminer la taille des zones d'agrégation. Toutefois, nous devons nous conformer à quelques considérations sur le choix de la taille. Ainsi la taille doit être :

1. suffisamment grande pour éliminer le bruit blanc ;

2. suffisamment fine pour capturer la structure spatiale du phénomène ;
3. suffisamment fine pour disposer d'un nombre acceptable de points pour l'apprentissage.

Cela se traduit numériquement pour le cas d'application sur la distance angulaire, respectivement sur ces trois aspects comme suit :

1. La variance du bruit blanc est 3 fois plus grande que celle de la structure spatiale (la variance du signal à l'échelle de la structure spatiale). Après intégration, nous souhaitons que l'écart type du bruit blanc soit 3 fois moindre que celui de la structure (écart-type du signal à l'échelle de la structure spatiale) et si nous devons l'éliminer complètement, nous proposons de le diminuer 10 fois. Il faut donc au moins 80 bâtiments par zone à raison d'une densité pour la zone d'étude 1300 bâtiments par kilomètre carré. Ce qui donne une zone de de taille minimale de 250 m x 250 m.

2. Pour capturer finement cette structure, il est impératif de disposer d'au moins 1 point tous les 400 m.

3. Avec des zones de 400 m de côté, on obtient un nombre de zones égal à $6943 * 4374/(400 * 400) = 190$ zones, ce qui semble suffisant pour la phase d'apprentissage.

S'il faut donc choisir des zones (en fonction de la distance euclidienne) uniquement à partir de l'autocorrélation de DA , il semble intéressant de prendre des zones de 300 à 400 m de coté. A l'issue de cela, on définit une structure spatiale de taille de 400m de côté environ en se basant sur un schéma d'agrégation des données.

6.2.1.2 Constitution de la structure spatiale et de ses voisinages

Afin d'identifier les critères nous permettant de définir des nouvelles structures spatiales (i.e. zones agréées), nous nous sommes inspirés des travaux de recherche antérieurs (citer les travaux). Ceux-ci nous ont permis d'identifier trois indicateurs qui nous semble pertinents à savoir **la régularité, la similarité et la proximité**. Ainsi, nous souhaitons dresser un découpage basé sur : l'alignement, la régularité, la proximité, l'orientation des bâtiments, etc.

En respectant le seuil défini par le covariogramme, la première étape consiste en la formation des groupements des bâtiments suivant l'intersection des routes pour former d'une structure spatiale dite **îlot**.

La deuxième étape consiste à répartir le groupe des bâtiments dans le même ilot, en des sous-groupes en rattachant à chaque batiment du groupe à sa route la plus proche de sorte à constituer un groupe de bâtiments généralement homogènes du fait qu'ils soient situés tous le long d'une même route, ce qui implique souvent que ces bâtiments ont des alignements, des orientations et même parfois des formes similaires. Cette structure de sous-groupe est dite **sous-îlot**, c'est l'appartenance à une route ou rue en terme de distance. La structure spatiale finale retenue est celle composée de la combinaison d'îlot et de sous-îlot. Nous l'appelons formellement la structure **îlot/sous-îlot**. Cette dernière a priori,

est destinée à contenir des bâtiments homogènes vis-à-vis de la forme, de la taille et de leur disposition géographique comme illustré sur la figure 6.11.

En pratique, nous nous procurons le réseau routier de la zone d'étude. Sur ce réseau routier, nous générons un graphe plan. Un réseau plan se caractérise par des arcs qui ne s'intercoupent pas. Cela implique de disposer d'un réseau où chaque tronçon entre deux sommets du réseau a un identifiant propre comme illustré sur la figure 6.9.

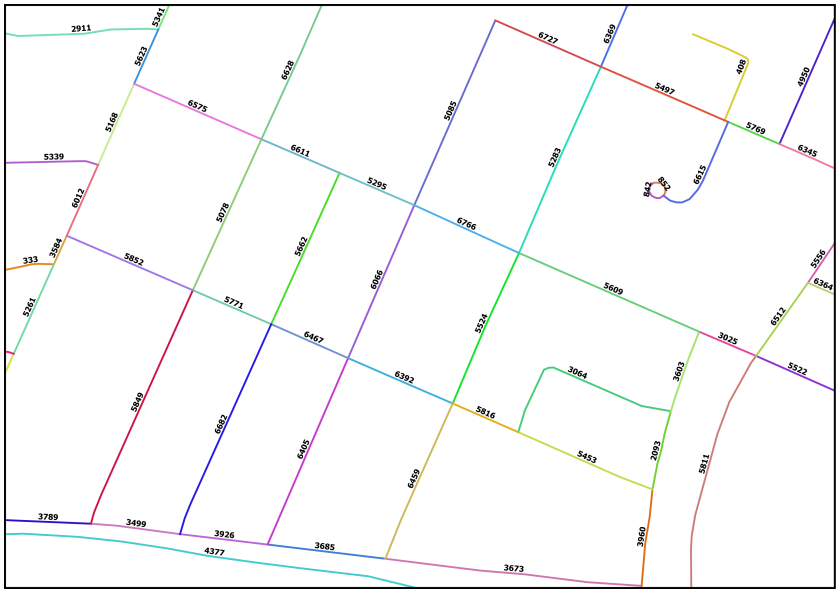


FIGURE 6.9 – Illustration du réseau routier plan. Chaque arc porte un numéro unique et ne coupe à aucun autre arc.

Sur le réseau plan obtenu, nous réalisons une extraction des faces du réseau. Ici, les faces correspondant à nos îlots. La figure 6.10 illustre quatre faces coloriées différemment, conduisant à la définition de quatre îlots. Chaque îlot doit porter un numéro unique.

Sur les faces extraites du réseau, nous réalisons une intersection avec la couche de nos bâtiments afin d'attribuer à chacun de nos bâtiments le numéro de la face dans laquelle il se trouve. Sur cette face, nous recherchons l'arc dont le bâtiment est le plus proche (cela peut-être un arc interne qui ne sépare pas deux faces). Nous finissons par attribuer à notre batiment, le numéro de l'arc le plus proche dans la face où il se trouve. Ainsi, le batiment sera référencé par un numéro composé du numéro de la face dans laquelle il se trouve et celui de l'arrête la plus proche de cette même face.

Sur la figure 6.11, nous représentons les sous-îlots (ici la rue) avec la même couleur que celle des bâtiments qui lui sont rattachés. Nous remarquons que des bâtiments peuvent être rattachés au même tronçon de rue tout en se situant dans des îlots voisins.

Suivant cette structure spatiale dite **îlot/sous-îlot**, nous définissons la relation de voisinage entre un bâtiment et un autre bâtiment comme suit. Pour un batiment donné :

- le 1er voisinage contient tous les bâtiments se trouvant dans le même sous-îlot et de

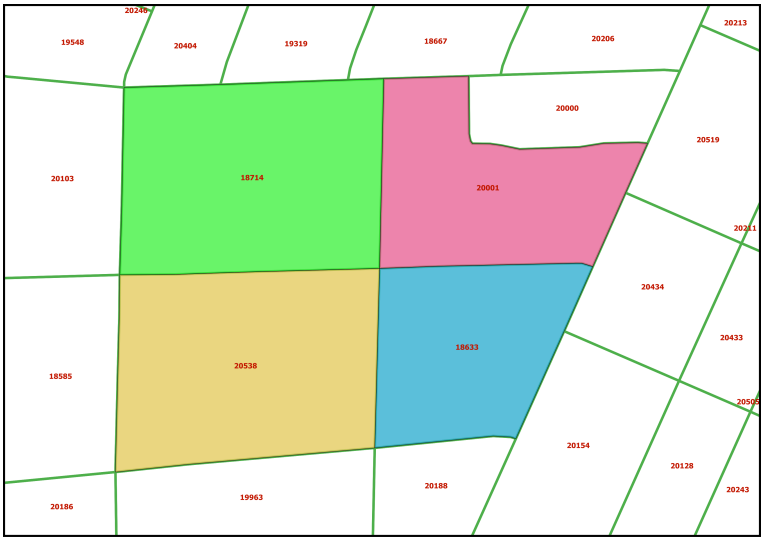


FIGURE 6.10 – Illustration d'un exemple des faces générées à partir du réseau routier donnant lieu à des îlots.

FIGURE 6.11 – Représentation de la structure spatiale à l'échelle de la **sous-îlot** avec une répartition des bâtiments en fonction de leur appartenance à la **sous-îlot**. Les numéros en bleu correspondent aux identifiants de **l'îlot**, les identifiants en noir aux identifiants de rue, définissant les **sous-îlot**. Nous colorions en noir les bâtiments et les sous-îlots non traités dans l'exemple.

le même îlot ;

- le second voisinage est formé par les bâtiments se situant sur un autre sous-îlot mais toujours dans le même îlot ;
- 3ème voisinage est celui formé par un groupe de bâtiments se trouvant de l'autre coté de sa rue (c'est-à-dire, dans le même sous-îlot mais d'îlots voisins) ;
- Le 4ème voisinage du batiment en question serait celui constitué par les bâtiments qui appartiennent à un sous-îlot différent que celui du batiment mais mais toujours dans un îlot voisin ;
- le 5ème voisinage du batiment donné se construit autour des bâtiments qui sont dans un îlot non voisin et différent que celui du batiment en question.

Sur la figure 6.12, nous représentons les différents voisinages autour d'un batiment donné en partant des bâtiments au sein de l'îlot vers les bâtiments se trouvant dans l'îlot adjacent (ou opposé) jusqu'aux bâtiments de l'îlot non-opposé.

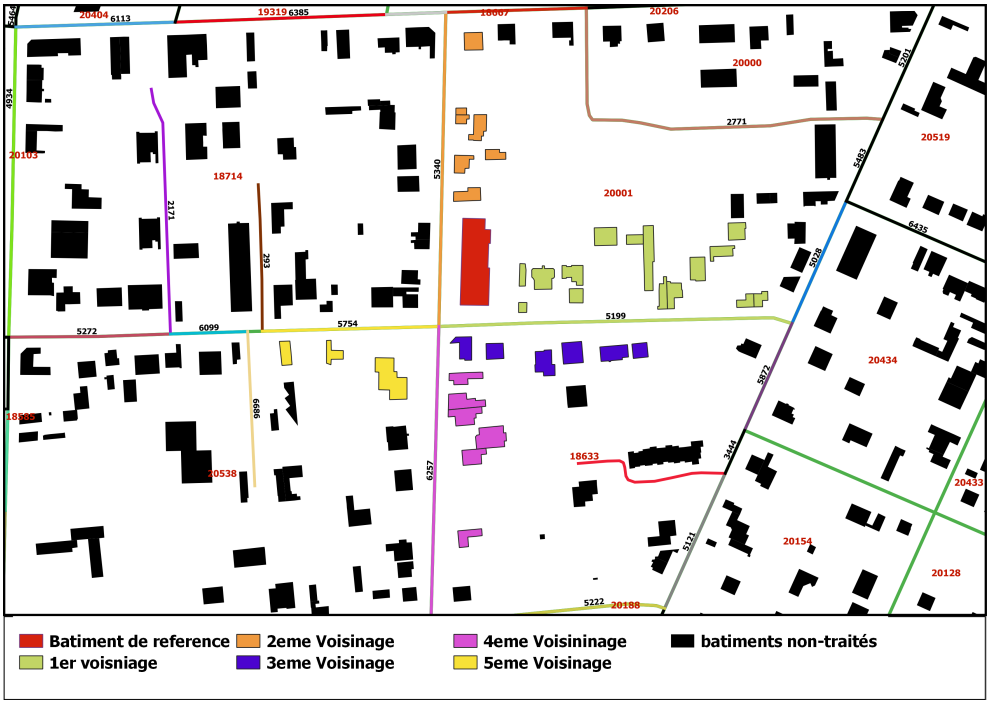


FIGURE 6.12 – Représentation des différents voisinages d'un batiment (en rouge) avec ses voisins bâtiments à l'échelle de la structure**îlot/sous-îlot**. Ici les bâtiments constituants un agrégat dans la structure en question seraient les bâtiments se situant dans le 1er voisinage. L'ordre de voisinage est donné par la légende de la figure.

Ainsi, nous constituons un agrégat de bâtiments à partir d'un ensemble de bâtiments se situant dans un même sous-îlot au sein d'une même îlot. C'est sur cet agrégat de bâtiments que nous souhaitons justifier l'existence d'une auto-corrélation.

A l'issue de la construction du schéma de la structure spatiale et de la définition du voisinage, nous recherchons s'il existe une auto-corrélation des valeurs de la variable dépendante (le cas ici, la distance angulaire) pour pouvoir ré-étudier le modèle de régression à l'échelle de l'agrégat (en moyennant la variable dépendante sur l'ensemble des valeurs des

bâtiments se trouvant dans une même structure spatiale, c'est-à-dire la structure **îlot/sous-îlot**).

6.2.1.3 Évaluation de l'auto-corrélation spatiale dans le voisinage défini par la structure spatiale

Après avoir défini un découpage en forme de structure spatiale **îlot/sous-îlot**, nous définissons une manière de calculer une valeur d'auto-corrélation spatiale sur l'ensemble des voisins d'un batiment. Cela consiste d'abord à définir une matrice de poids entre un bâtiment et son voisin bâtiment appartenant à un des 5 voisinages définis auparavant. Puis nous calculons une valeur globale indiquant la pertinence d'une d'auto-corrélation spatiale (l'indice de Moran).

Pour l'ensemble des 5 voisinages définis plus haut, nous attribuons un poids dégressif au fur et à mesure que le batiment voisin se trouve à un niveau de voisinage d'ordre décroissant (du 5ème voisinage au 1er voisinage) pour pouvoir calculer l'indice de Moran. Ainsi, pour chaque couple de bâtiments $b1$ et $b2$, on définit le poids associé au couple (b1,b2) de la manière suivante :

- Les deux bâtiments se trouvent dans deux îlots différents et non-voisins (5ème voisinage), leur poids est fixé à $poids = 0$
- Les deux bâtiments se trouvent dans deux îlots différents et voisins mais des sous-îlots non-opposés, leur poids est fixé à $poids = 0,25$
- Les deux bâtiments se trouvent dans deux îlots différents et voisins et encore des sous-îlots opposés (de part et d'autre sur la même rue), leur poids est fixé à $poids = 0,5$
- Les deux bâtiments se trouvent dans le même îlot mais des sous-îlots différents, leur poids est fixé à $poids = 0,75$
- Les deux bâtiments se trouvent dans le même îlot et dans le même sous-îlot, leur poids est fixé à $poids = 1$

A l'issue de ces cinq configurations, pour un batiment b_i donné, ses voisins sont l'ensemble des bâtiments b_j tels que le poids w_{ij} soit différent de 0. Nous calculons donc la valeur de l'auto-corrélation spatiale d'une variable dépendante y (e.g. distance angulaire) en utilisant ces relations de voisinage et ces poids à travers l'indice de Moran sur le voisinage d'un batiment :

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{6.17}$$

avec $z_i = (x_i - \bar{x})$; $z_j = (x_j - \bar{x})$; $w_{i,j} = poids$ pour le couple de bâtiments (b_i, b_j) ; n : nombre total des bâtiments ; S_0 : la somme total des poids ;

Ainsi l'indice de Moran **I** est calculé comme suit :

- pour chaque batiment b_i, nous calculons la somme du produit entre z_i, les z_j de ses voisins et leurs poids $w_{i,j}$ que l'on note (1) ;
- pour chaque batiment b_i, nous calculons la somme z_i^2 que l'on note (2) ;

- nous réalisons le rapport des sommes des termes (1) et (2) que l'on note (3) ;
- nous calculons l'indice de Moran en multipliant (3) par $\frac{n}{S_0}$.

Si la valeur de l'indice de Moran est égale à 0, on affirme qu'il n'existe pas d'auto-corrélation entre les valeurs de la variable. Plus la valeur tend vers -1 (respectivement 1), plus nous affirme qu'il a une auto-corrélation spatiale négative (respectivement positive). Toutefois, les limites entre ces valeurs ne sont pas bien nettes car elle dépendent de la sensibilité du phénomène étudié. C'est pourquoi, nous nous demandons si la valeur de l'auto-corrélation spatiale est statistiquement significative. En effet nous nous demandons dans quelle mesure l'agencement spatial de la variable dépendante n'est pas dû au simple fait du hasard. Pour confirmer ou rejeter l'hypothèse de l'existence d'auto-corrélation spatiale, nous avons suivi la démarche suivante :

- nous réalisons des permutations des valeurs associées aux agrégats, en conservant leurs positions. On effectue cette opération 29 fois. On obtient 29 jeux de données.
- pour chaque permutation, nous calculons l'indice de Moran. Cela permet d'obtenir 29 nouvelles valeurs d'indice de Moran.
- nous calculons le *Z-score* : nous comparons le vrai indice de Moran à la distribution des indices de Moran des permutations. En supposant l'hypothèse de base (H_0) que la valeur de l'indice de Moran obtenu serait le produit du hasard et pourrait provenir de la distribution des indices de Moran issus des permutations, nous calculons ce score statistique *Z-score* qui mesure à quel point notre indice de Moran s'éloigne de la distribution des indices Moran aléatoires.

6.2.1.4 Méthode de régression à l'échelle de la structure spatiale

Après avoir confirmé l'existence d'une auto-corrélation spatiale, nous construisons un nouveau modèle de régression. Ce dernier porte à l'échelle de la structure spatiale dite *îlot/sous-îlot*. Ici l'individu ou l'entité spatiale sur laquelle porte la régression est un agrégat des bâtiments appartenant au même îlot et au même sous-îlot. Pour chaque structure *îlot/sous-îlot*, nous calculons une valeur agrégée de la variable dépendante. La valeur agrégée est donc obtenue en calculant la moyenne des valeurs prises par les bâtiments pour une variable dépendante sur l'ensemble des bâtiments composant la structure *îlot/sous-îlot*. Bien que nous considérons dans ce modèle de régression à l'échelle du l'*îlot/sous-îlot* pour les mêmes variables dépendantes, les variables explicatives connaissent l'ajout de quelques variables explicatives exprimant une description globale de l'agrégat. Nous souhaitons combiner des descripteurs sur un ensemble d'indicateurs moyennés et des descripteurs portant sur l'ensemble de la structure spatiale.

Ces variables explicatives décrivant la structure spatiale *îlot/sous-îlot* sont : la régularité de l'agrégat, similarité de l'agrégat, et la proximité de l'agrégat (espacement entre les bâtiments).

- **régularité (reg)** : cet indicateur mesure l'orientation globale des bâtiments dans la structure spatiale en comptant le nombre de fois où il y a eu un changement majeur d'orientation d'un bâtiment à un autre. Nous notons un changement majeur de régularité quand l'angle d'orientation du PPR d'un polygone est 2 fois supérieur (ou

inférieur) à celui du polygone qui le précède. Nous incrémentons alors le nombre de régularité. Une valeur faible de ce nombre traduit une régularité des bâtiments au sein de la structure spatiale ;

- **similarité (sim)** : cet indicateur mesure à quel degré les bâtiments dans la structure spatiale diffèrent de taille t de surface. Il mesure le nombre total majeur de changement de surface dans la structure spatiale. Nous notons un changement majeur de similarité quand la surface d'un polygone est 2 fois supérieur (ou inférieur) à celui du polygone qui le précède. Nous incrémentons alors le nombre de similarité. Ce nombre est d'autant plus faible que les bâtiments dans la structure spatiale se ressemblent davantage en terme de surface.
- **proximité (proxim)** : cet indicateur mesure le nombre de sous-groupes dans la sous-îlot en fonction d'un changement majeur d'espacement entre les bâtiments. Nous notons un changement majeur d'espacement quand le périmètre d'un polygone est 2 fois supérieur (ou inférieur) à celui du polygone qui le précède. Nous incrémentons alors le nombre de proximité. Ce nombre est d'autant faible que la structure spatiale est homogène.

Une fois que le découpage en sous-îlot s'est avéré pertinent après avoir prouvé l'existence d'une auto-corrélation spatiale, nous appliquons notre méthode de régression linéaire sur cette nouvelle structure donnant un agrégeant des bâtiments. La mesure prédite, sera une valeur moyenne pour un groupe de bâtiments au dépens d'une prédiction locale à l'échelle du bâtiment. Ainsi, nous reprenons toute la démarche de réalisation, d'analyse et de validation du modèle de régression décrite dans la section 6.1.1.2. Avec une validation croisée couplée d'un bootstrap sur les données, nous raffinons notre modèle de régression en lui appliquant à une pénalisation LASSO afin d'obtenir que les variables plus importantes.

Pour pouvoir retenir l'échelle idéale sur laquelle, il est judicieux d'établir le modèle final de régression pour chacune des variables dépendantes, nous comparons les résultats sur les performances des modèles de régression obtenus à trois échelles à savoir, à l'échelle du batiment, à l'échelle de la structure *îlot/sous-îlot* et enfin à l'échelle de la structure *îlot*. Bien évidemment, pour justifier l'existence d'autocorrélation positive à l'échelle de la structure *îlot*, la même démarche ayant abouti à justifier l'existence d'une auto-corrélation spatiale (auto-corrélation spatiale positive) à l'échelle de la structure *îlot/sous-îlot*, a été entreprise.

6.2.2 Résultats d'estimation des indicateurs extrinsèques à partir des indicateurs intrinsèques à l'échelle de la structure spatiale

6.2.2.1 Résultats sur la pertinence de l'autocorrélation spatiale

A la suite du découpage effectué sur notre jeu de données de bâtiments selon les tailles préconisées par notre analyse lors de la proposition du schéma d'agrégation des bâtiments, nous avons obtenu 190 zones de taille environ 400 m sur lesquelles nous agrégeons environ 80 bâtiments en moyenne dans chaque structure *îlot/sous-îlot*. C'est sur ces agrégats que nous détaillons les résultats obtenus pour le calcul de l'indice de Moran puis pour l'établissement du modèle de régression.

Concernant l'auto-corrélation spatiale, comme précédemment expliqué, nous calculons l'indice de Moran en utilisant les poids déterminés sur les voisinages des bâtiments à

l'échelle de la structure spatiale *îlot/sous-îlot*. Les résultats donnent une valeur d'indice d'autocorrélation spatiale égale à **0.021**. A première vue, ce score semble être faible pour affirmer l'existence de l'autocorrélation spatiale. Il est plus rigoureux d'évaluer la significativité de l'indice de Moran à travers le calcul du **Z-score**.

Pour cela, nous représentons sur la figure 6.13, une comparaison faite entre valeur la valeur vrai de l'indice de Moran obtenue par notre étude d'auto-corrélation spatiale et la distribution de valeurs des indices de Moran issues des permutations aléatoires faites sur les données de base. Le principe consiste donc à exprimer à quel degré notre valeur d'indice de Moran s'éloigne à la valeur moyenne des indices de Morans issus des permutations. Le score d'éloignement est donné par la valeur du *Z-score*.

Ainsi, sur la figure 6.13, nous observons que la valeur vrai de l'indice de Moran est très éloignée de toutes les valeurs possibles obtenues par des permutations. En réalité, la valeur du *Z-score* étant égale à 22, nous estimons que notre valeur d'indice de Moran est **22** fois plus éloignée de la moyenne des indices de Moran issus des permutations. Cela confirme fortement l'existence d'une auto-corrélation spatiale et rejette par conséquent l'hypothèse disant que la valeur de l'indice de Moran calculée sur notre de jeu de données pourrait être obtenue par un pur hasard, avec une *p-value*, si tant est que l'on veuille bien lui prêter un sens, de l'ordre de 10^{-106}.

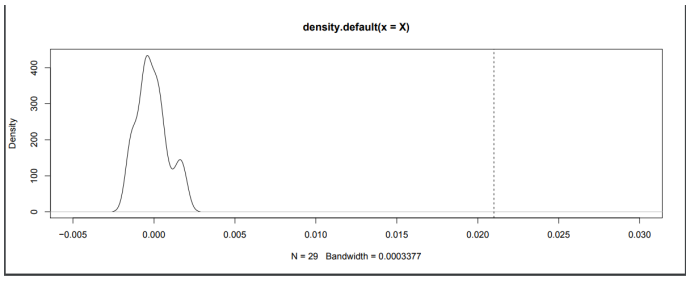


FIGURE 6.13 – Illustration du Z-score entre la valeur de Moran observée et la distribution des indices de Moran issus des permutations

La significativité forte de l'existence d'une auto-corrélation spatiale justifie notre logique d'agrégation des bâtiments à l'échelle de la structure spatiale avec laquelle il est possible de rehausser le taux de variance expliquée d'une régression faite sur une valeur moyennée (d'une variable dépendante) sur les valeurs des bâtiments au sein de la structure spatiale.

Pour pousser encore plus loin notre raisonnement sur le rehaussement de la performance de notre modèle de régression, nous avons cherché à savoir le potentiel de la variance expliquée (c'est la quantité maximale de variance qu'on peut espérer expliquer en travaillant à l'échelle du **îlot/sous-îlot**). Ainsi, selon le théorème de la variance totale, la variance totale du phénomène (le cas de la distance angulaire) est partagée entre la variance inter-classe et la variance intra-classe. Ici, la classe correspond à l'agrégat obtenu sur la structure *îlot/sous-îlot*. Nous calculons chacune des variances :

$$\mathrm{Var}(intra) = 0.895$$

et

$$\mathrm{Var}(inter) = 2.191$$

Un rapide calcul nous donne alors la part de variance explicable R^2 :

$$R^2 = \frac{\mathrm{Var}(inter)}{\mathrm{Var}(inter) + \mathrm{Var}(intra)}$$

$$R^2 = \frac{\mathrm{Var}(inter)}{\mathrm{Var}(inter) + \mathrm{Var}(intra)} = \frac{1}{(1 + \frac{\mathrm{Var}(intra)}{\mathrm{Var}(inter)})} = 1/(1 + 1/0.7099) = 1/2.409 = 71\%$$

Cela correspond pour le cas de la distance angulaire, **71%** de variance **explicable** par une régression parfaite sur la structure **îlot/sous-îlot**.

Si nous prenons en compte également l'indice de Moran (avec un **Z-score de l'ordre de 22**) en plus de ce résultat, nous disposons d'arguments solides pour affirmer que le découpage en **îlot/sous-îlot** est pertinent pour mieux prédire l'indicateur extrinsèque (ici la distance angulaire) à partir des indicateurs intrinsèques à l'échelle de **îlot/sous-îlot**.

6.2.2.2 Résultats sur le modèle de régression à l'échelle de la structure spatiale

Pour ce qui suit, nous détaillons les résultats obtenus avec la régression LASSO munie d'une une validation croisée à **2-fold** pour chaque indicateur extrinsèque à l'échelle de la structure **îlot/sous-îlot**. Nous relevons également les variables explicatives les plus pertinentes que contiendra le modèle de régression afin d'estimer une valeur sur un indicateur extrinsèque :

- **Distance angulaire** à l'échelle de la structure **îlot/sous-îlot** :
  Dans la table 6.12, nous représentons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée valant à : $R^2 = 27.65\%$

Coeffs. a	$\hat{a}$
a_{elg}	0.055548
a_{cpc}	−0.0548543
a_{agl}	−0.030746
a_{out}	0.0304047
a_{lmn}	−0.0165161

TABLE 6.12 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance angulaire du modèle de régression LASSO à l'échelle de la structure **îlot/sous-îlot**.

Ce qui donne comme variables explicatives retenues : **élongation**, **compacité**, **angle-droit**, **outlier**, **longueur-minimale** et donnant ainsi l'équation de prédiction du modèle

final de régression :

$$\widehat{Y} = 6.39.10^{-5} + 0.055x_{elg} - 0.054x_{cpx} - 0.03x_{agl} + 0.03x_{out} - 0.0165x_{lmn} \quad (6.18)$$

- **Distance radiale** à l'échelle de la structure **îlot/sous-îlot** :
  Dans la table 6.13, nous illustrons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée qui est de : $R^2 = 15.38\%$

Coeffs. a	$\hat{a}$
a_{out}	0.09773
a_{cpc}	−0.00655
a_{lme}	0.00368

TABLE 6.13 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance radiale du modèle de régression LASSO à l'échelle de la structure *îlot/sous-îlot*.

Ce qui donne comme variables explicatives retenues : *outlier* et *rapport-long-minimale-maximale*, et donnant ainsi l'équation de prédiction du modèle final de régression :

$$\widehat{Y} = 0.001168 + 0.00368x_{lme} + 0.09773x_{out} - 0.006558x_{cpc} \quad (6.19)$$

- **Distance Hausdorff** à l'échelle de la structure **îlot/sous-îlot** :
  Dans la table 6.14, nous affichons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée valant à : $R^2 = 5.9\%$

Coeffs. a	$\hat{a}$
a_{ragl}	−0.127034
a_{out}	0.141859
a_{elg}	0.051839

TABLE 6.14 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance de Hausdorff du modèle de régression LASSO à l'échelle de la structure **îlot/sous-îlot**.

Ce qui donne comme variables explicatives retenues : **angle-droit**, **outlier** et **élongation**, et donnant ainsi l'équation de prédiction du modèle final de régression :

$$\widehat{Y} = 0.010479 + 0.141859x_{out} - 0.12703x_{ragl} + 0.051839x_{elg} \quad (6.20)$$

- **Distance surfacique** à l'échelle de la structure **îlot/sous-îlot** :

Dans la table 6.15, nous représentons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée valant à : $R^2 = 3.32\%$

Coeffs. a	$\hat{a}$
a_{cpc}	-0.15538
a_{ragl}	-0.13868
a_{lmn}	-0.0317790
a_{orient}	-0.002093

TABLE 6.15 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance surfacique du modèle de régression LASSO à l'échelle de la structure **îlot/sous-îlot**.

Ce qui donne comme variables explicatives retenues : **compacité**, **angle-droit**, **longueur-minimale**, *orientation-sous-îlot*, *outlier*, et donnant ainsi l'équation de prédiction du modèle final de régression :

$$\widehat{Y} = 9.79.10^{-5} - 0.155x_{cpc} - 0.138x_{ragl} - 0.031x_{lmn} - 0.002x_{orient} \tag{6.21}$$

Enfin pour pouvoir mener une comparaison des performances des modèles de régression à différentes échelles, nous détaillons les résultats du modèle de régression obtenues à l'échelle de la structure dite **îlot**. Cela sous-entend de constituer l'agrégat à partir de l'ensemble des bâtiments se trouvant dans un même îlot.

- Distance angulaire à l'échelle de la structure **îlot** :

  Dans la table 6.16, nous représentons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée valant à : $R^2 = 41.46\%$

Coeffs. a	$\hat{a}$
a_{cvx}	$-1.470091.10^{-1}$
a_{elg}	$9.279215.10^{-2}$
a_{lmx}	$-7.339749.10^{-2}$
a_{out}	$5.686135e-02$
a_{ragl}	$-4.076994.10^{-2}$

TABLE 6.16 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance angulaire du modèle de régression LASSO à l'échelle de la structure **îlot**.

- Ce qui donne comme variables explicatives retenues : **convexité**, **élongation**, **longueur-minimale**, **outlier**, **angle-droit** et donnant ainsi l'équation de prédiction du modèle final de régression :

$$\widehat{Y} = 6.135.10^{-5} - 0.147x_{cvx} + 0.0927x_{elg} - 0.0733x_{lmx} + 0.0556x_{out} - 0.0407x_{ragl} \tag{6.22}$$

- **Distance radiale** à l'échelle de la structure **îlot** :

  Dans la table 6.17, nous illustrons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée qui est de : $R^2 = 38.05\%$

Coeffs. a	$\hat{a}$
a_{cvx}	-0.67743512
a_{cpc}	-0.28824186
a_{rec}	0.25688640
a_{ragl}	-0.01440350

TABLE 6.17 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance radiale du modèle de régression LASSO à l'échelle de la structure *îlot*.

Ce qui donne comme variables explicatives retenues : *convexité* et *compacité*, *rectangulaire* et *angle-droit*, et donnant ainsi l'équation de prédiction du modèle final de régression :

$$\widehat{Y} = 4.405.10^{-4} - 0.6774x_{cvx} - 0.2882x_{cpc} + 0.2568x_{rec} - 0.0144x_{ragl} \tag{6.23}$$

- Distance Hausdorff à l'échelle de la structure **îlot** :

  Dans la table 6.18, nous affichons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée valant à : $R^2 = 11.70\%$

Coeffs. a	$\hat{a}$
a_{ragl}	-0.61715142
a_{elg}	0.38416152
a_{lmx}	0.12070549
a_{out}	0.08030291
a_{orient}	-0.04420362

TABLE 6.18 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance de Hausdorff du modèle de régression LASSO à l'échelle de la structure **îlot**.

Ce qui donne comme variables explicatives retenues : **angle-droit**, **elongation** et **longueur-maximale**, **outlier** et **orientation**, et donnant ainsi l'équation de prédiction du modèle final de régression :

$$\widehat{Y} = 0.004393 - 0.61715x_{ragl} + 0.3841x_{elg} + 0.1207x_{lmx} + 0.0803x_{out} - 0.0442x_{orient} \quad (6.24)$$

- Distance surfacique à l'échelle de la structure **îlot** :
  Dans la table 6.19, nous représentons les coefficients des variables explicatives retenues par le modèle de régression LASSO avec un taux de variance expliquée valant à : $R^2 = 6.05\%$

Coeffs. a	$\hat{a}$
a_{cvx}	0.525415604
a_{cpc}	−0.454214907
a_{agl}	−0.184027684
a_{elg}	0.056616973
a_{lme}	0.038946102

TABLE 6.19 – Illustration des valeurs estimées des coefficients (a) des variables explicatives pour la distance surfacique du modèle de régression LASSO à l'échelle de la structure *îlot*

Ce qui donne comme variables explicatives retenues : **convexité**, **compacité**, **angle-droit**, **elongation**, **rapport-longueur-minimale-longueur-maximale**, et donnant ainsi l'équation de prédiction du modèle final de régression :

$$\widehat{Y} = 0.0001759 + 0.525cvx - 0.4542x_{cpx} - 0.184x_{agl} + 0.056x_{elg} + 0.0389x_{lme} \quad (6.25)$$

En guise de comparaison, nous récapitulons les taux des variances expliquées des modèles de régression obtenus aux trois échelles (bâtiment, îlot/sous-îlot et îlot) dans le tableau 6.20 suivant donnant les taux variances :

$R^2 en\%/echelle$	$echelle\ du\ batiment$	$sous-ilot$	$ilot$
R^2_{da}	25.43%	27.65%	41.46%
R^2_{dr}	14.2%	15.38%	38.05%
R^2_{ds}	4%	3.32%	6.05%
R^2_{dh}	3%	5.9%	11.70%

TABLE 6.20 – comparaison des scores de performances du modèle de régression obtenu à trois échelles de bâtiments.

Dans un premier temps, nous constatons que pour les indicateurs de forme (distance angulaire et distance radiale), la performance de prédiction du modèle de régression s'améliore très légèrement en passant d'une régression à l'échelle du bâtiment vers l'agrégat à l'échelle de la structure spatiale dite **îlot/sous-îlot**. Cependant, à l'évolution vers un agrégat sur une échelle d'îlot, nous observons une nette amélioration de la performance de régression pour les indicateurs de forme. Il est donc clair que l'agrégation tendant jusqu'à l'échelle **îlot** fournirait apport de performance de prédiction sur une valeur d'indicateur

extrinsèque basée sur les distances de forme.

Par ailleurs, pour les indicateurs de position (distance surfacique et la distance de Hausdorff), bien que l'agrégation des bâtiments aux échelles supérieures a contribué à améliorer modestement la qualité de prédiction du modèle de régression, cette amélioration n'est toujours pas notable pour suggérer fortement d'adopter un modèle de régression basé sur une structure spatiale plutôt qu'à son échelle de base (échelle du bâtiment). Néanmoins, la faiblesse de taux de variance expliquée pour les indicateurs de forme étant dûe de base à une faible significativité d'établissement de corrélation entre ces indicateurs de forme et les indicateurs intrinsèques, l'idée d'agrégation demeure légitime en soi dans une quête d'amélioration de performance du modèle de régression.

En somme, le choix de changement d'échelle en passant par une justification d'existence d'une auto-corrélation spatiale, semble bien adéquat dans le but de mieux expliquer la variance des variables dépendantes afin de mieux estimer la qualité extrinsèque.

6.3 Relaxation du problème et traitement

Bien que le problème d'estimation d'une qualité spatiale soit de nature quantitative si bien que sa modélisation par une régression semble bien justifiée, nous avons toutefois remarqué que la régression ne donne pas des résultats très satisfaisants. Nous transformons le problème d'apprentissage en recherchant à détecter une qualité plutôt qu'à estimer numériquement cette qualité. Nous nous contentons d'identifier un bâtiment mal saisi au lieu de rechercher combien ce dernier s'écarte de son homologue d'une base de référence en termes de précision spatiale.

Dans l'étude que nous souhaitons mettre en place, la classification tente à répondre une question qualitative à travers un problème binaire (classe positive ou classe négative). Il faut donc rechercher une frontière entre les deux classes. En l'occurrence dans notre cas, nous recherchons la limite de détection entre ce que nous définissons comme **une donnée du bâti de qualité insuffisante** et **une donnée du bâti de bonne qualité**. Dans une régression, la fonction de régression estime ou prédit une valeur sur la variable étudiée tandis qu'en classification, elle estime plutôt une probabilité d'appartenance à la classe positive à partir de l'information issue des descripteurs (variables explicatives).

La principale différence entre les deux méthodes porte sur la notion de seuil de décision. Pour une régression, la variable à estimer (e.i. la qualité spatiale) étant déterminée de manière continue, un modèle de régression laisse le choix à l'utilisateur du seuil de séparation entre qualité insuffisante et bonne qualité tandis qu'une classification fixe ce seuil *a priori*. A partir de ce seuil, on détermine indirectement une *probabilité a priori* pour les deux classes.

De plus, la qualité de détection d'une classification est relative à la *probabilité a priori*. Si par exemple, la *probabilité a priori* de la classe positive est égale à 90% et que la précision de détection (proportion d'objets correctement classifiés) est de 90%, cela signifie que l'algorithme classe correctement 90% des objets, suivant cette mesure-là il n'est pas meilleur qu'un algorithme qui classerait tous les objets dans la classe positive. La détection n'est donc pas particulièrement performante en soit étant donnée la *probabilité a priori*. Sur ce

cas, la détection devrait être devrait être significativement supérieure à 90%. Cela traduit que la précision de détection (accuracy en anglais) n'est qu'un indicateur grossier qui seul ne permet pas une bonne évaluation d'une méthode de classification.

Nous pouvons comparer les performances de la classification et de la régression, soit en fixant un seuil d'acceptabilité dans la régression (et en l'évaluant comme une classification), ou en effectuant une classification avec plusieurs seuils pour se rapprocher d'une régression.

6.3.1 Méthode de classification caractérisant la qualité des bâtiments

Notre démarche de classification commence par fixer un seuil pour séparer ce qu'on appellera *données de bonne qualité* et *données de qualité insuffisante*. Une connaissance *a priori* de la qualité des bâtiments peut être utile pour fixer ce seuil, de sorte que les deux classes cibles soient à peu près équilibrées. A défaut de cette analyse, nous nous proposons de prendre la médiane comme seuil de séparation sur la variable d'étude. Ce qui donne une *probabilité a priori* égale à 0.5 pour la classe positive (resp. pour la classe négative) avec A = *bâtiment estimé de qualité insuffisante par le classifieur* et V = *bâtiment de de qualité insuffisante en réalité*. Nous attribuons la valeur **1** à toutes les valeurs inférieures à la médiane et **0** à celles supérieures à la médiane. C'est-à-dire que nous considérons comme *positifs* les bâtiments de qualité inférieure à la qualité médiane, et négatifs les autres. Nous nous plaçons ainsi dans une perspective de détection des bâtiments de qualité insuffisante.

Par la suite, le classifieur, nous donne une valeur de *probabilité à posteriori* pour chaque bâtiment, notée $P(V|descripteurs)$**.** Pour transformer la probabilité en une fonction de décision **F**, nous nous fixons un seuil t et nous décidons que $A = 1$ *si* $P(V|descripteurs) \geqslant t$**, sinon** $A = 0$. Ainsi, pour chaque seuil choisi, on peut calculer une spécificité et une sensibilité sur le jeu de données d'entrainement comme décrit dans la section 3.4.2.1. La courbe ROC représente la spécificité et la sensibilité de l'algorithme pour les différents seuils de t, et permet de calculer un indicateur de performance globale sur la classification dit **AUC**.

Pour parvenir à calculer les **probabilités a posteriori**, nous implémentons un algorithme basé sur la méthode *des forêts aléatoires*. Comme déjà mentionné, l'algorithme de *forets aléatoires* est une combinaison de plusieurs *arbres de décision*. Cette combinaison permet de réduire l'instabilité liée aux arbres individuels face à de légers changements dans les données (qui peuvent intervenir dans la phase de bootstrap) produisant des arbres très différents. En particulier, les changements affectant les nœuds proches de la racine affectent beaucoup l'arbre produit. On dit que les arbres produisent des estimateurs de variance élevée. Afin de remédier à ce problème de variance, les *forêts aléatoires* utilisent, la moyenne des probabilités obtenues sur l'ensemble des arbres de décision (Breiman, 2001; Bostrom, 2007).

Tout d'abord, dans un algorithme de *forêt aléatoire*, nous tirons au hasard dans la base d'apprentissage B échantillons avec remise $zi, i = 1, \ldots, B$ (chaque échantillon ayant n points). Pour chaque échantillon i , nous tire au hasard et sans remise, un sous-ensemble q ($q = \sqrt{p}$) parmi les p descripteurs. Avec ces q descripteurs, on construit un arbre $G_i(x)$ de sorte qu'un nœud découpe l'espace des descripteurs par un plan séparateur, en minimisant une fonction d'impureté traduisant l'hétérogénéité des étiquettes des données situées de part et d'autre de la séparation (mesure de l'entropie). Dans le cas où un nœud ne contient pas des données totalement homogène, nous souhaitons qu'il existe une classe (positive ou

négative) très majoritaire. Ainsi, les arbres sont moins corrélés car ils sont construits sur des ensembles différents de descripteurs et sur des échantillons différents. Dans la méthode de *forets aléatoires*, nous utilisons des arbres qui ne sont pas très profonds, petits donc moins performants, mais l'agrégation compense ces faiblesses, ce qui minimise l'entropie et le sur-apprentissage.

Ensuite, après avoir construit une collection de T arbres aléatoires, l'inférence sur une nouvelle donnée X^{n+1} est réalisée en calculant les probabilités conditionnelles sur chaque arbre, puis en moyennant les probabilités obtenues.

Après cela, pour mesurer la performance du modèle sans nécessiter de jeu de données de test, nous calculons l'erreur **out-of-bag (OOB)**. En effet, chaque arbre de décision est construit sur **63%** du jeu d'entrainement, le reste servant à tester la prédiction de l'arbre. On mesure donc l'erreur *OOB* comme étant le taux d'erreur empirique mesuré sur l'échantillon *OOB*. Selon la littérature, il est admis que cette procédure est suffisante pour évaluer la performance de la classification, ou pour comparer plusieurs modèles opérant sur des sous-ensembles différents de descripteurs (Meneroux, 2019; Genuer et al., 2010).

Enfin l'algorithme des *forets aléatoires* nous offre la possibilité d'estimer l'importance relative des descripteurs dans le processus de classification. L'importance d'une variable X_i donnée est calculée en fonction du gain en homogénéité des étiquettes des sous-arbres à chaque coupe dans laquelle elle intervient. En général, on prend le même critère d'homogénéité que celui qui a été utilisé pour la construction de l'arbre (Gregorutti, 2015). Cette mesure d'importance est particulièrement utile dans notre cas, où nous ignorons *a priori* quels descripteurs vont être informatifs dans le processus de décision, et où on souhaite pouvoir sélectionner un nombre restreint de descripteurs importants pour pouvoir mener une interprétation du modèle de classification.

Après l'implémentation de l'algorithme des *forêts aléatoires*, nous recherchons un point de fonctionnement sur la courbe ROC qui réaliserait le meilleur compromis entre sensibilité et spécificité, noté t. Bien que la valeur de l'AUC suffise pour déterminer la qualité de la classification, il est aussi essentiel de repérer le point optimal sur la courbe ROC afin de décrire les résultats sur la variable cible en termes de spécificité et sensibilité. Le point optimal s'obtient par la minimisation de la distance au point (0, 1). Le point (0, 1) représente le cas idéal où la sensibilité et la spécificité valent toutes deux 1. On cherche le point de fonctionnement (t) pour lequel le point correspondant sur la courbe ROC est le plus proche du point (0, 1). En utilisant la distance euclidienne, ce critère s'écrit :

$$Y = (1 - S_e(t))^2 + (1 - S_p(t))^2 \quad minimal \tag{6.26}$$

où S_e=sensibilité et S_p=spécificité

Nous remarque que ce critère donne la même importance aux fonctions S_e et S_p, donc aux deux erreurs de classement.

6.3.2 Résultats de l'inférence avec le modèle de classification

La méthode de classification présenté ci-dessus a été appliqué sur le même jeu de données OSM à l'échelle du bâtiment. Les mêmes quatre distances ont été utilisées pour détecter un batiment de qualité insuffisante. Les résultats sont présentés d'abord sur les

distances de forme (distance radiale et distance angulaire), puis sur les distances de positon (distance de Hausdorff et distance surfacique).

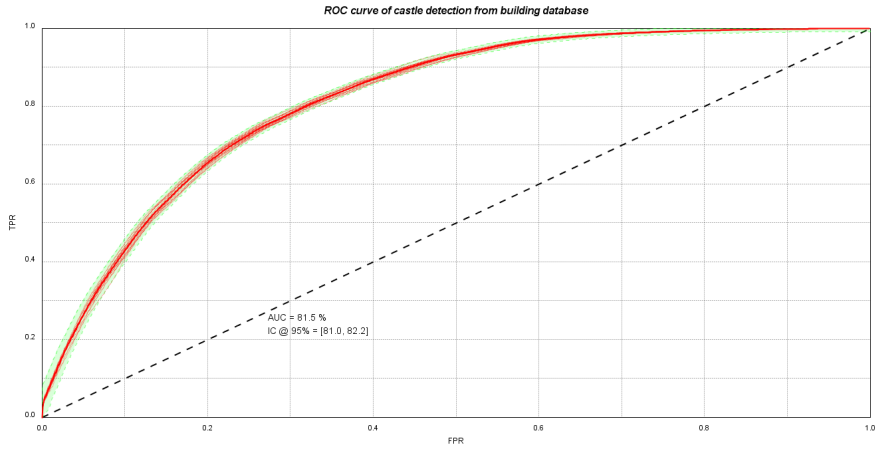


FIGURE 6.14 – courbe ROC : cas de la distance radiale

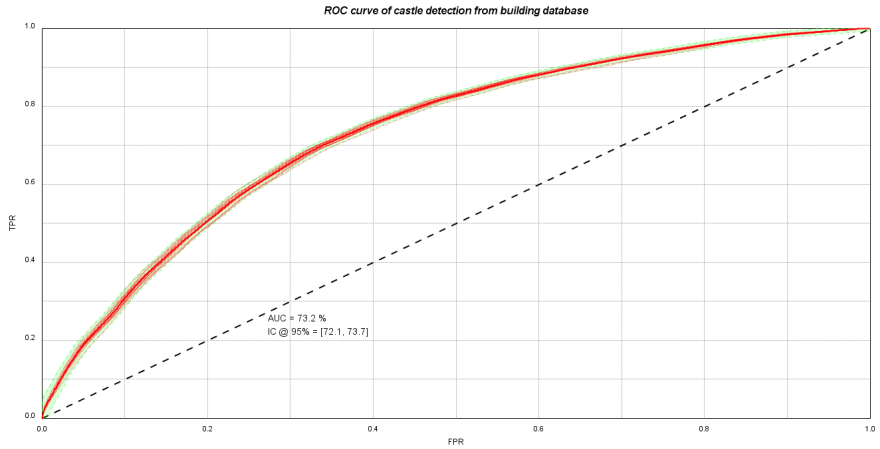


FIGURE 6.15 – courbe ROC : cas de la distance angulaire

Sur les indicateurs de forme, la tendance générale illustre que les courbes ROC convergent assez *vite* vers l'axe $y = 1$ quand FPR tend vers **1**. Cela indique que le classifieur peut récupérer quasiment toutes les données labellisées 1 à condition de faire un tri a posteriori pour retirer environ 50% de données faux positifs.

Pour les indicateurs de forme :

Distance radiale : sur la courbe de ROC de la figure 6.14, nous identifions le point optimal ayant 70% pour **spécificité** et de **80%** pour **sensibilité**. Avec une valeur d'AUC égale à **81.5%**, nous affirmons que le classifieur est capable de retrouver **80%** des bâtiments, parmi tous les bâtiments de qualité insuffisante. Selon la précision, l'estimateur se trompe à **27%** sur la détection, en y rajoutant des bâtiments de bonne qualité. Pour notre cas d'étude,

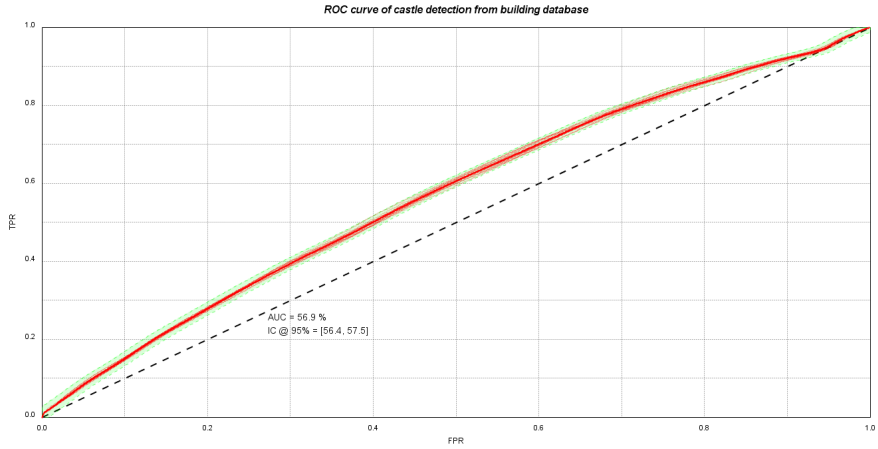


FIGURE 6.16 – courbe ROC : cas de la distance surfacique

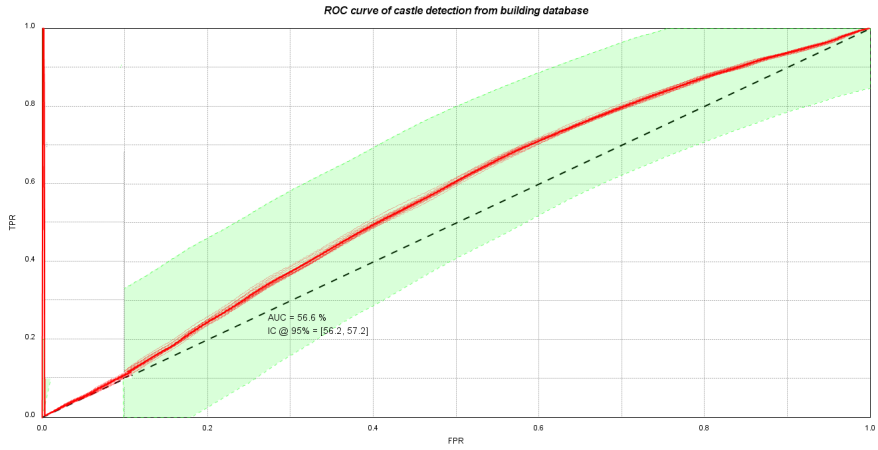


FIGURE 6.17 – courbe ROC : cas de la distance Hausdorff

si nous souhaitons adopter une méthode possédant un bon rappel, quitte à accepter la présence de faux positifs, nous pouvons affirmer, que le classifieur donne des très bons résultats sur la distance radiale.

Distance angulaire : sur la courbe de ROC de la distance angulaire (voir figure 6.15, le point optimal nous donne **50%** de **spécificité** avec **82%** de **sensibilité**. Bien que le classifieur détecte une très grande proportion des bâtiments de qualité insuffisante, il conserve un fort taux de faux positifs à **50%**. Cela nécessite un tri a posteriori sur les bâtiments détectés pour la distance angulaire. Ce résultat sur la distance angulaire peut être considéré comme satisfaisant à condition d'effecteur un tri a posteriori.

Quant aux indicateurs de position, les résultats ne semblent pas être encourageants :

Distance surfacique : sur la figure 6.16 au point de fonctionnement optimal, l'estimateur détecte à **70%** (Sensibilité) des bâtiments de qualité insuffisante. Pour cette valeur, sa Spécificité est de **40%**, si bien qu'il classe plus de la moitié des bâtiments bons comme mau-

vais. De tels résultats ne sont pas suffisants pour envisager une exploitation opérationnelle. Ces résultats sur la distance surfacique semblent pas satisfaisants vu qu'elles nécessitent des nombreux efforts a posteriori.

Distance Hausdorff : sur la figure 6.17 nous observons **40%** de rappel et **25%** de spécificité avec une valeur d'AUC égale à **56%**,ce qui nous indique que sur cet indicateur, le classifieur n'est pas du tout bon. Nous estimons donc que la classification semble ne pas apporter des résultats intéressants.

Chapitre 7

Transférabilité de l'algorithme de classification de la qualité des bâtiments

Sommaire

Dans ce chapitre, l'objectif principal est de mener une étude de transférabilité du modèle de classification sur d'autres zones d'études. On cherche à étudier quelle est la performance de la prédiction du modèle de classification, entrainé sur la zone de départ, sur une autre zone. Cela permet de discuter dans quelle mesure le modèle se généralise afin de tirer des conclusions sur la généricité ou la particularité de notre modèle d'inférence de la qualité extrinsèque à partir de la qualité intrinsèque.

7.1 Méthodologie de la Transférabilité

7.1.1 Étude des caractéristiques des zones d'études

7.1.1.1 Zone d'étude : GERS

Notre étude de transférabilité a porté sur une première nouvelle zone extraite du Département **GERS** en région Occitanie. La zone d'étude extraite s'étend sur une fenêtre rectangulaire de 72.4×42.8 km couvrant une surface de 3138 km^2.

Le jeu de données extrait se concentre principalement autour du centre des communes du département et de la ville Auch, et se disperse largement entre les communes à savoir au centre la commune **Marambat**, à l'ouest la commune de **Nagaro**, au nord-ouest la commune **Eause**, au sud-ouest la commune **Plaisance** au nord-est les communes **Fleurance, Lectoure et Saint-Clair** et enfin au sud-est la ville **Auch** comme illustré sur la figure 7.2.

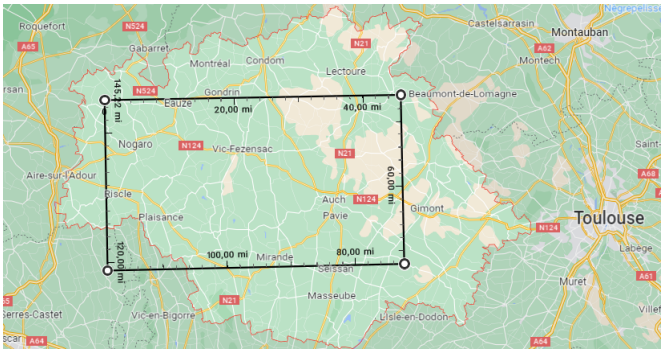


FIGURE 7.1 – Représentation de la zone d'étude du GERS

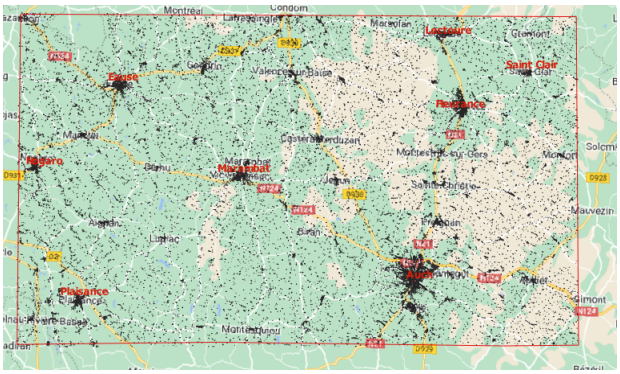


FIGURE 7.2 – Répartition spatiale du jeu de données extrait dans la zone d'étude extraite du GERS(32) (encadrée en rouge). Le nuage des points correspond à une forte concentration des polygones du bâti sur le centre des communes

De manière générale, ces communes sont qualifiées de communes rurales peu ou très peu denses avec des zones urbanisées autour de 3.6% sauf la ville d'Auch qualifiée comme étant une ville dense ou intermédiaire avec un taux de surface urbanisée d'environ 14.9% au sens de la grille communale de densité de l'Insee [1] [2] [3] [4] [5].

En s'inspirant du rapport de présentation sur le plan local d'urbanisme de **Miélan** [6], nous dressons une typologie générale de l'habitat au sein des communes du GERS. La morphologie générale de l'habitat se caractérise par une diffusion de l'habitat sur le territoire agricole mettant en place un éparpillement des constructions sur le territoire (sous forme de hameau ou d'habitat isolé) qui tend à être bâni de plus en plus. Plus précisément le

1. Selon le zonage des communes rurales et urbaines publié en novembre 2020, en application de la nouvelle définition de la ruralité validée le 14 novembre 2020 en comité interministériel des ruralités.

2. « Typologie urbain / rural » [archive], sur www.observatoire-des-territoires.gouv.fr (consulté le 27 mars 2021).

3. « Comprendre la grille de densité » [archive], sur www.observatoire-des-territoires.gouv.fr (consulté le 27 mars 2021).

4. « Comprendre la grille de densité » [archive], sur www.observatoire-des-territoires.gouv.fr (consulté le 27 mars 2021).

5. « CORINE Land Cover (CLC) - Répartition des superficies en 15 postes d'occupation des sols (métropole). » [archive], sur le site des données et études statistiques [archive] du ministère de la Transition écologique. (consulté le 14 avril 2021)

6. `https://www.gers.gouv.fr/content/download/23811/171130/file/`

rapport distingue quatre principales typologies des communes rurales que l'on retrouve sur le GERS à savoir typologie centre bourg, typologie lotissement, typologie d'organisation particulière et enfin la typologie d'habitat diffus.

Étant la typologie dominante au centre des communes, la typologie **centre bourg** se caractérise par une forte liaison entre l'espace public et les constructions et une accolement d'un bâti dense et implanté en alignement sur la rue (constitution d'un front continu). Les hauteurs de bâti sont très majoritairement en R+1 ou R+2 (Rez de Chaussée surmonté d'un ou deux étages). Cette typologie marque des logiques d'organisation dont les bâtiments servent de limites s'adossant entre au bâtiment voisin comme illustré sur la figure 7.3.

FIGURE 7.3 – Illustration de la typologie du bâti en bourg. Source : **https://www.gers.gouv.fr/content/download/23811/171130/file/**

Quant à la typologie dite **lotissement**, le bâti est dressé sous forme de lotissement donnant des parcelles régulières souvent en forme carrée et dont les constructions s'implantent au cœur de la parcelle, souvent sans mitoyenneté. Il existe des retraits par rapport à la voie et les maisons sont de nature pavillonnaire (voir la figure 7.4).

FIGURE 7.4 – Illustration de la typologie du bâti en lotissement. Source : **https://www.gers.gouv.fr/content/download/23811/171130/file/**

Concernant la typologie du bâti dite **d'organisation particulière**, le bâti s'implante en limite de la voirie. Il est essentiellement composé des bâtiments allongés spécifiques

des corps de ferme, occupant des parcelles de formes et de tailles hétérogènes. La limite parcellaire est bien délimitée par une haie ou un espace boisé comme illustré sur la figure 7.5.

FIGURE 7.5 – Illustration de la typologie du bâti en organisation particulière. Source : **https://www.gers.gouv.fr/content/download/23811/171130/file/**

Enfin la morphologie du bâti dans les communes du GERS est caractérisée par une typologie **d'habitat diffus**. Cette typologie favorise un étalement urbain impactant le milieu agricole. Le bâti s'implante sans rapport avec le tissu alentour. Il se situe en milieu du parcelle avec création d'une voie d'action privé. Cette typologie se caractérise par une heterogeneité des constructions comme illustré sur la figure 7.6.

FIGURE 7.6 – Illustration de la typologie du bâti en mode habitat diffus. Source : **https://www.gers.gouv.fr/content/download/23811/171130/file/**

En somme la zone d'étude extraite du GERS caractérise principalement par un milieu rural peu dense dont l 'espace urbanisée est très défavorisée par rapport à l'espace agricole. Le bâtiments ont davantage des formes allongés et sont espacées sauf au centres des communes où on observe un habitat dense et accolé. L'habitat pavillonnaire demeure est prépondérant. La zone d'étude répond donc à l'objectif de travailler sur une zone de morphologie d'habitat en mode rural afin de ressortir des caractéristiques propres en milieu rural et communes avec l'habitat en milieu rural observé dans la zone d'étude de base (département Val-de-Marne).

7.1.1.2 Zone d'étude : Édimbourg (Écosse)

Dans la perspective d'utiliser des données extraites sur la ville d'Édimbourg, nous analysons la morphologie urbaine de la ville afin de dégager ses caractéristiques, et ceci dans l'objectif de mieux expliquer les résultats de notre analyse de transférabilité.

FIGURE 7.7 – Bâti sur le zone d'étude de la ville d'Édimbourg (Écosse). On représente en noir, les bâtiments de la base de référence sur la zone d'étude.

La zone d'étude s'étend sur une superficie d'environ 9 km² (3,019km d'Est en Ouest, et 2,947km du Nord au Sud). Nous pouvons distinguer trois sous-zones marquée par une différence de typologie du bâti. Au Nord, la sous-zone appelée communément *New Town* (nouvelle ville) abrite des bâtiments modernes sous forme d'une succession de blocs composés d'immeubles de deux à trois étages. Ce type de bâti est marqué par des bâtiments de forme rectangulaire, respectant un alignement parfait sur le réseau routier. Un large espace vert sépare cette sous-zone en deux, et des espaces verts plus petits se trouvent près de certains blocs. C'est donc un bâti de type résidentiel moyennement dense se situant en zone urbaine.

Au centre se trouve la sous-zone appelée communément *Old Town* qui renferme un en-

FIGURE 7.8 – la sous-zone (appelé New town) au Nord de la ville d'Édimbourg (Écosse)

semble de bâtiments anciens de forme parfois irrégulière, et qui comprennent beaucoup de bâtiments administratifs, des écoles, des hôpitaux, des musées ainsi que des églises. C'est également là où se trouvent la plupart des commerces de la ville. C'est plutôt une zone commerciale avec un bâti assez dense marqué par des ruelles étroites. Les bâtiments sont très rapprochés sans respecter une disposition particulière. Cette sous-zone est séparée par la sous-zone *New Town* par un grand espace vert et un réseau de voies ferrées.

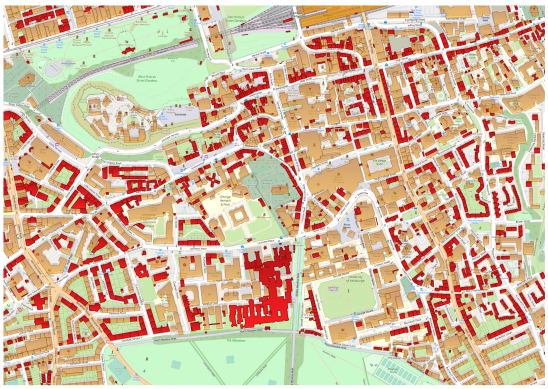


FIGURE 7.9 – la sous-zone (appelé Old town) au Centre de la ville d'Édimbourg (Écosse). En rouge, on représente les bâtiments de la base de référence, en marron les bâtiments de la base OSM.

Enfin, au Sud on retrouve une sous-zone mixte. En dessous de la sous-zone *Old Town*, dans le quartier appelé *Marchmont* se trouve des immeubles de quatre et cinq étages construits dans le style baronnial écossais [7]. La plus grande partie de la zone a été développée dans les années 1870 et 1880 et sa structure n'a guère changé depuis lors. Plus à droite, dans le quartier de *Bruntsfield*, on retrouve un paysage similaire, c'est-à-dire un quartier résidentiel avec des immeubles en forme de barres, très allongés suivant la direction de la rue. Ce n'est qu'au Sud du *Bruntsfield* qu'on retrouve un bâti de type résidentiel pavillonnaire avec par endroits des habitations de type villas. Cette typologie s'étend d'Ouest en Est jusqu'au quartier *Newington*. Le paysage est marqué par une forte présence d'espaces verts. Les bâtiments sont espacés et les alignements sont moins repérables.

Analyse des résultats d'appariement sur la zone d'étude Édimbourg

Dans une démarche de validation de l'appariement sur la zone d'étude extraite de la ville *Édimbourg*, capitale de l'Écosse au Royaume-Uni, nous procédons à une étape de vérification des résultats de l'appariement. Au fur et à mesure que nous manipulons le plugIn de validation, nous relevons les cas rencontrés qui permettent comprendre la logique de l'algorithme d'appariement sur ce jeu de données d'Édimbourg, et ce afin de dégager la spécificité de la zone d'étude en matière d'appariement. Nous avons utilisé les mêmes initialisations de masses de croyances ainsi que les mêmes fonctions de croyances vis-à-vis des critères d'appariement. Le seuil de sélection des candidats est toujours de 30 m. En

7. https ://en.wikipedia.org/wiki/Marchmont

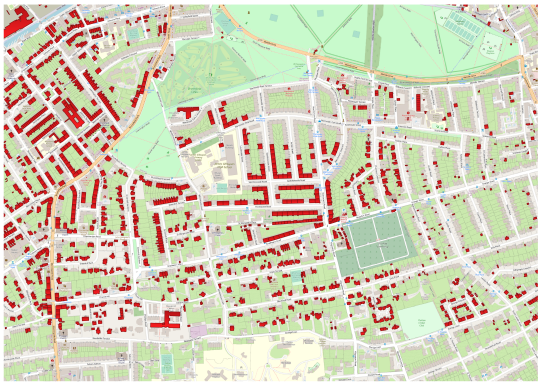


FIGURE 7.10 – la sous-zone situé au Sud. En rouge, on représente les bâtiments de la base de référence, en marron les bâtiments de la base OSM.

conservant ces seuils et ces fonctions de croyance, nous souhaitons mettre en évidence le caractère générique de l'algorithme d'appariement relativement au seuil de sélection et aux fonctions de croyance sur nos quatre indicateurs extrinsèques quelle que soit la typologie de la zone d'étude.

Si notre analyse sur l'appariement tente d'étudier les objets *appariés* et *non-appariés*, nous souhaitons également comprendre les cas d'indécision et apporter une explication à l'ensemble du spectre des résultats d'appariement. De manière globale, sur un jeu de données de 6131 objets de la base de référence sur la zone d'étude, 2565 objets s'apparient avec des objets de la base OSM, 1393 objets demeurent *non-appariés* tandis que 2123 objets donnent lieu à un verdict d'indécision, et seuls 20 objets ne reçoivent pas de candidats depuis la base OSM. Notre premier constat est que le taux d'appariement demeure faible avec beaucoup de cas d'indécision. Nous nous livrons alors à une vérification détaillée sur nos résultats d'appariement. Ainsi nous constituons un jeu de données de type bâti composé de 100 couples objets *Appariés* et 60 objets *Non-Appariés*. Avec un jeu de données de cette taille, nous pensons pouvoir estimer la performance de l'algorithme d'appariement en terme de Précision et de Rappel, et ceci en accord avec la méthode prescrivant la taille de l'échantillon représentatif décrite dans la sous-section 5.2.3.

Dans les cas d'indécision, on trouve des bâtiments saisis différemment dans les deux jeux de données. Comme premier exemple, un château ancien est saisi en un seul morceau dans la base de l'Ordonance Survey alors qu'il est découpé en plusieurs bâtiments parfois entre-coupés par des voies et des rues dans la base OSM comme illustré par la figure 7.11. Le bâtiment (dans la base de référence) s'étend sur une grande surface et est actuellement en construction (après vérification sur Street View). Sur la base de référence, ce bâtiment est constitué d'un seul polygone qui traverse parfois des voies tandis que ses candidats issus de la base OSM sont des polygones de tailles plus petites et semblent être des parties annexes du château. Ce bâtiment se situe à proximité de la place nommée *Lauriston*. On confirme bien qu'aucun objet de la base OSM lui ne ressemble dans son intégralité et de ce fait, ne peut lui être homologue bien que le polygone OSM situé en haut de la figure 7.11 semble être en partie fidèle à la forme du batiment en question. On estime que l'indécision provient du fait que les deux bases ont des spécifications assez différentes. Dans ce cas là, on souhaiterait que les objets de la base de référence soient découpés comme ceux de

la base OSM. Une autre perspective pourrait aussi être que l'algorithme d'appariement apparie un objet dans la base de référence avec plusieurs objets de la base OSM.

FIGURE 7.11 – Exemple illustrant un cas d'indécision dù probablement en partie (car il peut être dû aussi à une différence d 'actualité entre la base de reference et la base OSM) à une différence de spécification entre la base de référence et la base OSM dans l'hypothèse d'exclure . En rouge l'objet de la base de référence et en Vert les objets de la base OSM

Par ailleurs, sur un autre cas d'indécision, nous observons que l'objet de la base de référence est non seulement décalé mais semble correspondre à l'union de deux objets OSM voisins. Malgré ce possible rapprochement visuel qui s'opère, notre vérification valide l'indécision car les deux objets OSM sont comparés séparément avec l'objet de la base de référence (voir figure 7.12). Ici on pourrait proposer d'utiliser les relations spatiales (jointure des deux bâtiments) pour améliorer l'algorithme qui conserverait les relations entre les objets. Comme pour l'exemple précédent, on pourrait réduire le nombre de cas d'indécision en autorisant des liens d'appariement de type **1 :m**.

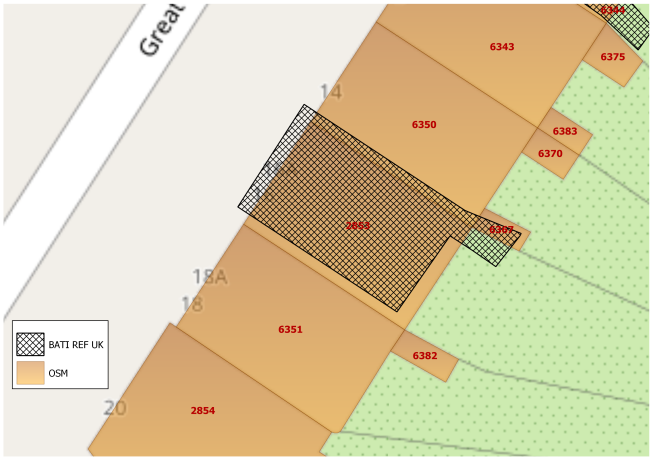


FIGURE 7.12 – Exemple illustrant un cas d'indécision dû probablement à une différence de spécification entre la base de référence et la base OSM. En hachure noir l'objet de la base de référence, et en marron les objets de la base OSM.

En somme, nous avons observé que dès lors qu'il existe un décalage (sur la base de référence) et un découpage (souvent sur les objets OSM), il en résulte un appariement indécis la plupart du temps, et parfois un *non-appariement* selon que la distance surfacique tend à reste entre $0,3$ et 0.5 (indécision), ou qu'elle soit supérieure à 0.5 (non-appariement). Cependant il apparaît parfois que des cas qui devraient donner lieu à un résultat d'indécision, donnent des résultats d'appariement. Cela dû est au fait que la distance Hausdorff a un poids plus important que la distance surfacique (à condition que la distance surfacique soit inférieure à 0.5). Ainsi l'algorithme apparie l'objet de la base de référence malgré des valeurs assez élevées des distances portant sur la forme (distance radiale et distance angulaire assez élevées par rapport à leurs seuils). De ce fait, on a dans la classe des objets appariés des appariements généralement justes, mais qui suivent deux motifs distincts : les premiers objets sont appariés avec des valeurs très faibles (qu'on peut qualifier de *parfaitement appariés*) pour toutes les distances, tandis que les seconds sont appariés malgré des valeurs importantes de la distance radiale et de la distance angulaire (qu'on pourrait qualifier *appariés de justesse*). Cela engendre un mélange d'objets appariés avec une forte hétérogénéité sur la distance radiale et sur la distance angulaire. En guise d'exemple des groupes d'objets correctement appariés, la figure 7.13 illustre deux objets de la base de référence, l'un étant parfaitement apparié et l'autre apparié de justesse. On recense une liste des valeurs sur les distances d'une part sur des objets :

- parfaitement appariés :

TABLE 7.1 – Quelques exemples illustratifs des appariements avec des valeurs très faibles des distances sur la forme

settings	*ds*	*dh*	*dr*	*da*
(1)	0.27	1.26	0.29	0.37
(2)	0.23	1.6	0.23	0.256
(3)	0.22	1.514	0.084	0.184
(4)	0.195	1.67	0.189	0.117
(5)	0.187	1.537	0.47	0.32

- appariés de justesse :

TABLE 7.2 – Quelques exemples illustratifs des appariements avec des valeurs assez grandes des distances sur la forme

settings	*ds*	*dh*	*dr*	*da*
(1)	0.3	2.782	1.546	0.66
(2)	0.3	3.26	1.58	0.59
(3)	0.41	6.88	1.626	0.59
(4)	0.39	7.38	2.02	0.33
(5)	0.357	7.86	2.3	0.68

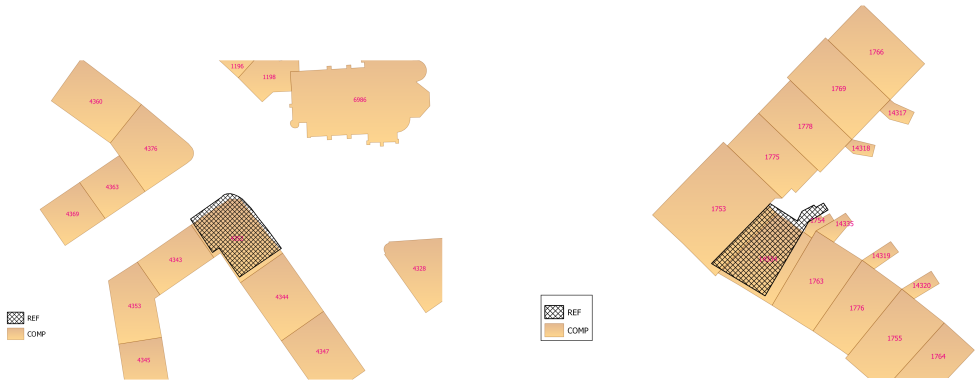


FIGURE 7.13 – Exemple illustrant deux objets de la base de référence correctement appariés. A gauche, on représente en rouge l'objet de la base de référence parfaitement apparié avec un objet OSM qui s'apparente le plus à l'objet de la référence. A droite un autre objet de la base de référence qui s'apparie difficilement avec un objet OSM avec une distance radiale et une distance angulaire assez hautes.

A l'issu de la vérification des objets *appariés* et des objets *non-appariés*, nous obtenons une Précision égale à **97%** avec un Rappel égal à **98,97%**. Sur ces scores, nous validons notre appariement tout en étant conscients que les valeurs des distances retenues (surtout sur les distances de formes comme mentionné sur le tableau) en tant qu'*étiquettes* pour le modèle d'apprentissage, présentent des grandes variances susceptibles d'atténuer les performances du modèle d'apprentissage qui en résulte.

7.1.2 Hypothèses

7.1.3 Démarche de l'étude de la transférabilité

Afin d'étudier la transférabilité, nous nous procurons trois jeu de données sur le bâti : un jeu de données sur notre 1ère zone d'étude sur Val-de-Marne (Ile-de-France, France), un second dans le Gers (Occitanie, France) et un troisième jeu de données dans Édimbourg (Écosse, Royaume-Uni). La démarche globale consiste à vérifier si le modèle entrainé sur une zone d'étude se transfère sur une autre zone d'étude. Par transférabilité, on entend par mesure de l'aptitude du modèle d'apprentissage à se généraliser à de nouvelles données test (données n'ayant pas servi à apprendre le modèle), c'est-à-dire, la capacité du modèle à parvenir à classer correctement les bâtiments issus d'une autre zone. Dans le cas où on apprend sur une partie de la zone et on valide sur une autre partie de la même zone d'étude, il s'agit d'une procédure de validation croisée, qui a pour objectif de faire ressortir la qualité de prédiction globale du jeu de données sur soi-même.

En fait, la transférabilité évalue la performance de l'algorithme sur de nouvelles données issues de jeux de données de caractéristiques différentes. Si la validation évalue les performances de l'algorithme sur des données non présentes dans le jeu d'entrainement, mais de caractéristiques a priori proches (car issues du même jeu de données). Par contre, la transférabilité mesure la capacité de l'algorithme a avoir de bonnes performances sur des données pouvant avoir des caractéristiques différentes (autre pays, autre organisation du bâti, etc.). Du fait de ces changements de caractéristiques, on s'attend à ce que l'algo-

rithme ait des performances moins bonnes sur ces nouvelles données que celles observées dans le processus de validation, mais on espère que cet écart de performance n'est pas trop important pour pouvoir tout de même utiliser l'algorithme sur des données variées.

Ainsi, nous concevons trois modèles différents, chaque modèle étant entrainé sur une partie des données d'une seule de trois zones d'étude. Puis nous évaluons les performances de ces trois modèles sur trois jeux de données de validation correspondant à ces trois zones, et nous comparons les mesures de performance pour ces 9 expériences.

Ainsi, nous découpons chaque jeu D_i en K_i folds, pour faire de la cross validation à K_i folds à l'intérieur de chaque jeu de données. Mais pour bien mener une étude de transférabilité, nous devons respecter certaines propriétés. Tout d'abord, nous souhaitons que les trois jeux d'entrainement aient la même taille dans toutes les expériences, pour que la comparaison des modèles soit plus homogène. Cela conduit à éventuellement choisir des valeurs de K différentes pour chaque jeu de données, et éventuellement à sous-échantillonner les jeux d'entrainement pour qu'ils aient tous le même taille.

Puis nous déterminons la taille nécessaire pour le jeu d'entrainement à l'aide de la formule ci-dessous, et cette taille détermine la valeur de K pour chaque jeu de données. De ce fait, pour s'assurer d'avoir le même nombre sur le jeu d'entrainement quelque soit la taille initiale du jeu tiré sur la zone d'étude, et garantir que le temps de calcul des termes diagonaux (validation croisée) et extra-diagonaux (jeu d'entrainement et jeu de validation tiré sur deux zones d'études différentes) soient sensiblement équivalents, il faut choisir un nombre maximal de données (pour l'entraînement) égal à :

$$NE^* = E\Big(\frac{N+1}{N+3} \times \min(n_1, n_2, ...n_N)\Big)$$

avec :
- $E(.)$ la partie entière
- N le nombre de datasets
- n_k le nombre de données du dataset k.

En particulier, on vérifie bien que NE^* tend vers $n-1$ avec n le nombre de données du plus petit dataset, quand le nombre de datasets tend vers l'infini (autrement dit, quand on a beaucoup de jeu de données, on peut se permettre de faire du *leave-one-out* sur les plus petits jeux de données). A l'inverse, quand N vaut 2 (cas limite minimal), on a $NE^* = E(3n/5) = 60\%$ du plus petit jeu de données, ce qui rapproche la méthode d'une validation croisée à 2 folds.

Dans notre cas d'exemple hypothétique, de trois jeux de données (N = 3) avec des efectifs différents pour les trois de données :
-$n_1 = 500$
-$n_2 = 554$
-$n_3 = 800$

donc : $n = \min(500, 554, 800) = 500$

d'où : $NE^* = E(4/6 * 500) = E(333.33...) = 333$ instances

Donc :

- Pour chaque jeu de donnée, l'évaluation des performances de l'algorithme entraîné sur ces données, avec validation sur les données du même jeu, se fait par validation croisée à K_i folds : on divise le dataset D_i en K_i groupes avec K_i= E(n_i/(n_i-NE*))+1.
  Ainsi dans notre exemple :
  - D1 est découpé en K1 = 3 groupes de taille 166, 167 et 167.
  - D2 est découpé en K2 = 3 groupes de taille environ 184, 185 et 185.
  - D3 est découpé en K3 = 2 groupes de taille 400.

  On effectue alors une validation croisée classique $K_i folds$ sur chaque jeu D_i, mais en s'assurant qu'à chaque itération, le jeu d'entraînement est *sous-échantillonné* pour ne jamais dépasser $NE* = 333$ données. Par construction, il est toujours possible d'avoir exactement NE* données dans chaque entraînement. Par exemple, sur D_1, à chaque fois qu'on choisit un groupe (ou fold) de validation, on a 2 groupes d'entraînement, soit dans le cas le pire 166+167=333 données. Sur D_2, à chaque groupe validé, on a 2 groupes d'entraînement, soit : $2*184 = 368$ données. On doit en retirer 368-333 = 35 à chaque entraînement. Sur D_3, pour chaque groupe de validation, on a 1 groupe d'entraînement, soit 400 données. On doit en retirer $400 - 333 = 67$ à chaque entraînement.

  L'évaluation de la performance s'effectuant par validation croisée à K_i folds, le calcul nécessite K_i itérations (i.e. la création de K_i forêts aléatoires).

  Par exemple, dans le cas de D_3, on a découpé le jeu en 2 morceaux : M_{31}, M_{32} de taille 400. On fait alors 2 itérations :
  - Itération 1 : entraînement sur M_{i1} et validation sur M_{i2} (avec M_{i1}= 333 données et M_{i2}=400)
  - Itération 2 : entraînement sur M_{i2} et validation sur M_{i1} (avec M_{i2}= 333 données et M_{i1}=400)

- pour l'évaluation des performances d'un algorithme entraîné sur les données d'un des jeux sur les données d'un autre jeu : on tire aléatoirement 333 données dans le dataset D_i pour entraîner le modèle et on valide sur toutes les données de D_j.

Au total, pour un nombre moyen de n instances par jeu, le calcul nécessite :
- $N * (N + 1)/2$ entraînements sur la diagonale (quand les données d'entraînement et les données de validation sont issues du même jeu).
- $N(N - 1)$ entraînements en dehors de la diagonale (quand les données d'entraînement et de validation sont issues de jeux différents).

Ainsi, à l'issue de ces considérations sur la taille du jeu d'entrainement et du jeu de validation, dans la matrice finale nous calculons pour chaque configuration la courbe de ROC reliant un ensemble des couples de valeurs (TPR ;FPR) pour divers seuils de séparation (seuil permettant de séparer les éléments étiquetés positifs des éléments étiquetés négatifs à partir des valeurs de probabilités fournies par la forêt aléatoire) et exprimons la performance globale du classifieur par la valeur de l'AUC. Nous précisons que la variable

dépendante utilisée est celle basée sur la distance radiale (car c'est celle qui donne les meilleurs résultats dans le chapitre précédent). Nous considérons donc la distance radiale comme variable d'étude pour la suite du chapitre.

Par ailleurs, sur chaque étape d'entrainement, nous relevons les variables explicatives les plus pertinentes afin de vérifier si les mêmes variables ont été sélectionnés pour entrainer le modèle de classification quelle que soit la zone d'entrainement ou si le modèle s'adapte aux données de la zone en sélectionnant un nouveau sous-ensemble des variables explicatives.

Enfin, dans un cadre de validation croisée (avec des données de validation issues du même jeu que les données d'entraînement), nous cherchons à déterminer la valeur la plus pertinente pour le seuil de séparation un seuil naturel a priori de séparation entre la classe positive (détecter une mauvaise qualité du bâti) et la classe négative (détecter une bonne qualité du bâti). Ce seuil est une valeur de distance radiale, en deçà de laquelle un batiment est considéré de bonne qualité (pour cette mesure). On calcule les valeurs de l'AUC pour différents seuils sur la distance radiale. On peut alors choisir ce seuil de manière à avoir la meilleure AUC possible. Avec ce seuil optimal, on examine dans le cas d'une validation sur un jeu d'une autre zone, la variation de la performance du classifieur en comparant la valeur d'AUC obtenue avec un seuil de égal à la médiane du jeu d'entrainement (choix qui permet de classer 50% des bâtiments du jeu d'entrainement comme de bonne qualité, et 50% comme de mauvaise qualité) avec celle obtenue avec le seuil optimal. Si le seuil optimal apporte aussi une amélioration pour des données de validation issues d'une autre zone, le jeu de validation est aussi partagé en deux classes suivant le seuil optimal. Ce dernier constitue désormais le seuil naturel de séparation aussi appelé seuil de binarisation.

7.1.4 Résultats sur l'étude de transférabilité

Dans le but d'analyser la transférabilité du modèle de classification, constituons trois jeux de données extraits aléatoirement sur nos trois zones d'études. Sur la zone d'étude du Val-de-Marne (94, Ile-de-France) (jeu dit *94*), on extrait un jeu comprenant 10530 instances, sur la seconde d'étude du GERS (jeu dit *GERS*), on tire 19068 instances et enfin sur la troisième zone d'étude à Édimbourg (Écosse) (jeu dit *UK*), nous extrayons 2558 objets du bâti. Ces nombres ont été choisis en tenant compte du nombre d'objets appariés sur les trois jeux de données même si les trois jeux de données ont des effectifs différents, les échantillons d'entrainement ont tous la même taille.

Dans un premier temps, nous réalisons l'apprentissage en fixant une probabilité a priori équitable pour la classe positive et la classe négative, car nous ne connaissons pas a priori la proportion de bâtiments de bonne qualité et de bâtiments de mauvaise qualité. Nous appliquons donc sur les trois jeux de données un seuil égal à la valeur médiane de la distance radiale. Nous analysons les résultats en utilisant des valeurs d'AUC pour évaluer la qualité de la classification. Ces valeurs d'AUC sont calculées pour chacun des croisements des jeu de données, les jeux de donnés étant croisés dans une matrice dont les diagonales correspondent correspondent à une opération de validation croisée sur un seul jeu de données et dont les éléments non diagonaux correspondent à l'utilisation d'un jeu de données pour l'entrainement et d'un autre pour l'évaluation. En guise de notation comme illustré sur la figure 7.14, nous désignons par **1** le jeu *94*, par **2**, le jeu GERS et enfin **3** par le jeu *UK* et pour chaque expérience **train** désigne le jeu d'entrainement et **valid** pour le jeu de

validation.

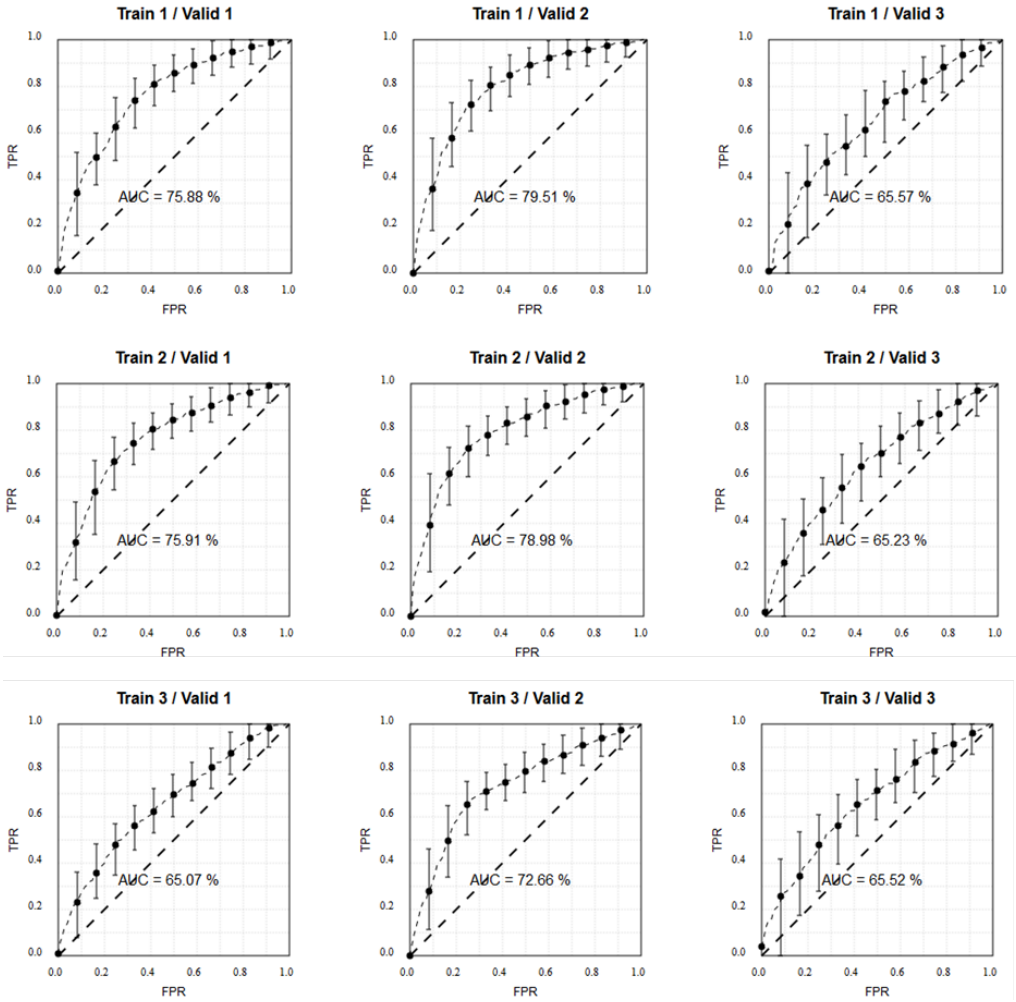


FIGURE 7.14 – Résultats de la transférabilité avec un seuil de séparation égal à la médiane du jeu d'entraînement. Les barres verticales sur la courbe de ROC indiquent un intervalle de confiance d'un écart type autour du point considéré. Une barre très grande traduit une instabilité des points constituant la courbe.

A première vue, on estime les modèles entraînés sur le jeu *94* conservent de bonnes performances quand ils sont évalués sur les deux autres jeux de données au moment du transfert du modèle de classification issu du *94* car on observe de manière globale qu'une bonne partie de la qualité de classification se conserve et se répercute lors de la validation sur les deux autres zones d'études. Plus en détail, nous observons que le modèle entrainé sur le *94* se transfert bien sur la zone du *GERS* (et même un peu mieux que le *94* sur lui-même) mais moins bien sur la zone d'étude du *UK* tandis que le modèle entrainé sur le *GERS* reste aussi assez généralisable sur le *94* et de la même façon que *94* a de moins bonnes performances sur le jeu du *UK*. Par contre le modèle entrainé sur le *UK* a une performance de classification faible et se comporte moins bien que les deux autres modèles. Bien que l'AUC prenne une valeur acceptable pour la validation sur le jeu du *GERS* (72%), le modèle entrainé sur le *UK* ne prédit pas mieux que les autres modèles

les données de la zone *UK*, et a de moins bonnes performances sur le jeu 94 et moins pour autant celles issues du *94*.

Si on tente d'apporter une spécification du modèle d'apprentissage en se basant uniquement sur les résultats de l'étude, nous estimons que le modèle issu du *94* est bien un modèle généraliste très bien entrainé sur un jeu diversifié. Le jeu *94* constitue à la fois un bon jeu d'entraînement (le cas **train1/train1**) et un bon jeu de test en cela qu'il se prête bien à la classification et permet d'évaluer de manière pertinente la qualité du modèle utilisé (ainsi il donne des résultats assez différents pour les modèles entraînés sur *94* et *GERS* d'une part, et pour le modèle entraîné sur UK d'autre part) (le cas **train2/valid1**). Le modèle produit sur le jeu du *GERS* est plutôt aussi un très bon jeu d'entraînement et de validation (le cas **train2/valid2**) mais un peu moins que le *94* en validation (le cas **train1/valid2**). Il semble être en général similaire à celui issu du *94*. Cela nous laisse penser que jeu du *GERS* dispose dispose de caractéristiques similaires à celui du *94* pour l'apprentissage, malgré la grande différence de leurs caractéristiques géographiques. Cette similarité eut s'expliquer par le fait que le jeu *94* comprend des zones d'habitat peu dense aux caractéristiques semblables à celles du jeu *GERS*.

Par contre, le modèle produit sur les données du *UK*, a les performances les plus basses quand les jeux *94* et *GERS* sont utilisés pour la validation, et le jeu UK lui-même n'est pas un très bon jeu de validation car les performances des trois modèles restent très basses sur ce jeu., on postule a priori que ce jeu a une grosse hétérogénéité sur ses données d'entrainement. Ce qui engendre à la fois une grosse difficulté à prédire un jeu de validation externe et une instabilité de prédiction sur ses mêmes données en validation croisée. Si ce modèle pourrait être considérer comme un bon modèle testeur, il n'est point un bon entraineur. Ces performances ne peuvent pas seulement s'expliquer par une structure différente du bâti dans ce jeu. Elles indiquent probablement que pour ce jeu la base de référence et la base qu'on cherche à évaluer (la base OSM) ont des spécifications trop différentes pour que nous puissions évaluer de manière pertinente la transférabilité de notre modèle.

Pour pouvoir comprendre les performances faibles du modèle d'apprentissage entraîné sur le jeu *UK*, nous investiguons les caractéristiques de la zone d'étude du jeu *UK*. Tout d'abord nous estimons que le modèle issu cette zone d'étude ne doit pas présenter un biais supérieur à celui des modèles générés sur les autres zones, car il a été produit avec le même nombre d'objets d'entrainement et les mêmes des variables explicatives. Sa qualité de classification est donc plutôt liée à la variance de l'estimateur lors de l'entrainement. Cette variance proviendrait comme déjà mentionné dans 7.1.1.2 lors de l'analyse des résultats de l'appariement sur le jeu de données *UK*, par des différences significatives entre les spécifications de la base de référence et les spécifications de saisie de la base OpenStreetMap. Ces différences engendrent des écarts parfois considérables entre les objets OSM et ceux de *l'Ordonance Survey*. Ces écarts très hétérogènes en terme d'amplitude et très nombreux affaiblissent l'apprentissage et le rendent également instable à chaque fois qu'on extrait un nouveau jeu de données pour l'entrainement. Bien que cette différence de spécification entre les données de référence et les données volontaires existe dans une certaine mesure pour tous les jeux de données, elle a été particulièrement marquée sur ce jeu de données. Sur les deux autres zones d'études, elle n'induit pas de répercutions majeurs sur la qualité du modèle de classification.

A l'issue de ces analyses, nous estimons que le modèle de classification que nous sou-

haitons (celui issu du *94*) transférer sur d'autres zones se généralise bien. Sa performance sur un jeu de données déterminé dépend des particularités liées aux données auxquelles nous l'appliquons, et notre étude a uniquement permis de mettre en évidence la qualité du transfert sur un autre jeu de données en France. Néanmoins, cette étude limitée confirme que ce modèle peut obtenir de bonnes performances sur de nouveaux jeux de données et sa qualité demeure assez bonne dans une grande envergure.

Ces résultats découlent de l'application d'un seuil égal à la distance médiane sur le jeu d'entraînement. Bien que ce seuil n'implique aucune décision a priori sur la qualité des objets de la base de validation, il n'offre pas nécessairement les meilleures conditions pour l'entraînement car qu'il peut y avoir des proportions différentes dans la base d'entrainement et dans le jeu de validation, et que le seuil choisi sur la base d'entrainement n'est pas nécessairement le meilleur pour le jeu de validation. Ceci peut pénaliser le modèle en terme d'apprentissage. En d'autres termes, nous souhaitons déterminer, pour un jeu donné, s'il existe un seuil qui s'approche du seuil naturel de séparation a priori des données qui donnerait un meilleur score pour une évaluation en validation croisée. Par la suite, nous appliquerons ce seuil aux validations du modèle sur les deux autres jeu de données. A ce stade, nous pensons abandonner le jeu *UK*, car nous affirmons pratiquement que les indicateurs produites avec ce jeu sont ininterprétables.

Ainsi, nous avons dressé des courbes mettant en évidence l'évolution de l'AUC en fonction de différents seuils pris entre 20% et 90% de la distribution du jeu d'entrainement (soit 65 valeurs de seuil au total). Pour chacune des deux zones d'études (94 et GERS), nous analysons la courbe de sorte à observer un point marquant un changement de régime d'AUC tout en restant le plus proche possible du seuil médian (50%) car les seuils extrêmes peuvent donner de meilleurs résultats, mais l'effectif de la classe minoritaire peut devenir faible, ce qui donne des résultats instables (très dépendants des données choisies pour l'entraînement).

Sur la figure 7.15, nous traçons l'évolution de l'AUC en fonction du seuil choisi sur le jeu *94*. On observe sur cette courbe qu'il existe une tendance globale croissante au fur et à mesure que l'on s'éloigne de la valeur centrale des seuils quand le seuil augmente. Sur cette courbe, nous proposons de choisir un seuil entre 60% et 65%. Au-delà de ces seuils, la courbe est très bruitée. Sur cette plage de seuils propices, nous retenons quelques exemples de seuils appliqués lors de validation sur les deux autres jeu de données pour retenir le seuil qui exprime au mieux la qualité de la classification sur l'ensemble des jeux de données et qui apporte une amélioration sur les scores précédents obtenus avec le seuil égal à 50%.

Sur la figure 7.16, nous illustrons les résultats de la classification sur la zone d'étude *94*, avec quatre seuils de séparation différents.

Par contre pour le cas du *GERS*, la fonction d'AUC possède une allure concave avec un maximum (voir figure 7.17). A des seuils faibles, l'AUC est bas, elle a une allure générale croissante avant de décroitre pour les seuils très élevés (supérieurs à 65%). Nous préconisons de choisir un seuil dans la plage de 62% à 65%.

Sur la figure 7.18, nous comparons les valeurs d'AUC pour des modèles entraînés sur le jeu *GERS* avec des seuils variables, pour justifier du choix du seuil le plus approprié

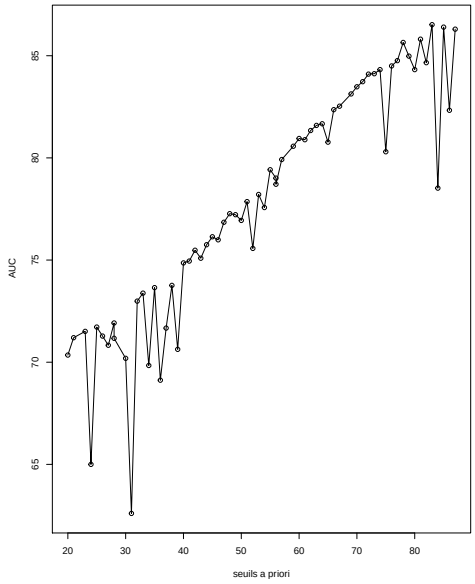


FIGURE 7.15 – l'évolution de l'AUC en fonction des seuils a priori le cas du *94*. Choisir un seuil de 65% revient à classer a priori 65% des bâtiments comme *bon* donnant une classification de mauvaise qualité.

pour ce jeu de données.

A travers de notre analyse nous avons montré que pour chaque zone, il existe un seuil optimal pour entraîner un modèle de classification et avec notre démarche, nous parvenons à le retrouver mais aussi à remarquer que ce seuil optimal améliore la qualité de la classification quand le modèle est appliqué sur la même zone ou sur une autre zone.

Ainsi, l'étude des seuil a permis d'augmenter la performance des modèles de classification créés en ce qui concerne l'AUC. Même s'il n'y a pas de valeur naturelle de seuil, elle peut être dictée par l'application, selon qu'on désire limiter les faux positifs, les faux négatifs, ou obtenir un compromis entre les deux.

L'analyse de la transférabilité passe aussi par une expérimentation autour des variables explicatives afin d'étudier d'un entrainement à un autre (en changeant la zone d'étude pour l'entrainement du modèle), les variables significatives sélectionnées. Nous relevons pour chaque entrainement sur un jeu données, les variables explicatives qui ont été utilisées par l'algorithme de classification par ordre d'importance. Voici les 6 variables les plus importantes pour le modèle entrainé sur :

- *94* :**SEGC-SEGL, OUTLIER, RECT, COMPACITE, CONVEXITE, SEG-LONG**
- *GERS* :**SEGC-SEGL, OUTLIER, RECT, CONVEXITE, COMPACITE, SEG-LONG**
- *UK* :**SEGC-SEGL, OUTLIER, CONVEXITE, RECT, SEG-LONG, COMPACITE**

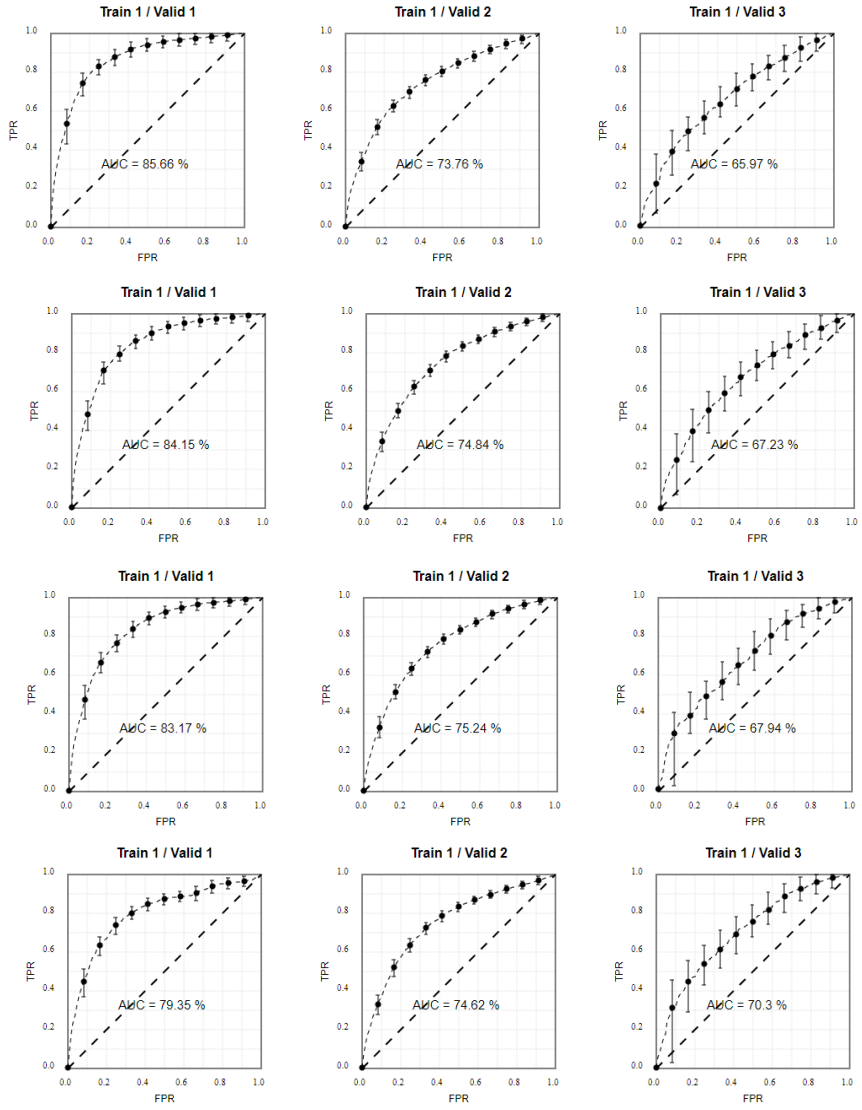


FIGURE 7.16 – Quelques exemples illustratifs de la qualité de classification pour différents seuils a priori. De haut en bas, on a les résultats de classification pour des seuils a priori respectivement à 70%, 65%, 62% et 60%. On observe bien que le seuil donnant les meilleures performances est égale à 60%.

Ainsi nous observons que l'algorithme de classification a choisi les mêmes variables explications pour l'ensemble des modèles d'apprentissages construits sur chacun de trois zones d'études. On pourrait dire que ces variables ont un intérêt qui ne se limite pas à un seul jeu de données, et qu'elles demeurent les plus importantes pour l'apprentissage. Cela renforce la pertinence de nos variables explicatives et surtout leur universalité sur les jeux de données et tend à confirmer l'idée que le modèle entraîné sur le jeu *94* peut être appliqué de manière pertinente sur de nouveaux jeux de données.

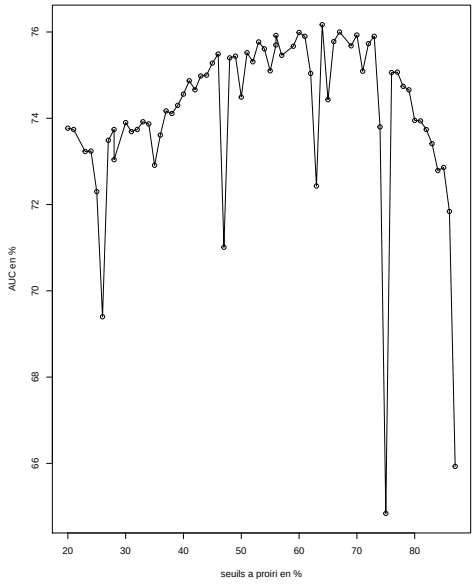


FIGURE 7.17 – l'évolution de l'AUC en fonction des seuils a priori pour le cas du *GERS*.

7.1.5 Application du modèle de classification sur Djibouti

7.1.5.1 Approches de fixation de seuil de qualité

A l'issu de la validation de la transférabilité du modèle de classification, nous effectuons une évaluation de la qualité extrinsèque à partir d'une implémentation des indicateurs intrinsèques faite sur un jeu de données d'OSM sur la ville de Djibouti en soumettant cette implémentation au modèle d'apprentissage. Ce dernier nous fournira pour chaque bâtiment du jeu de données de Djibouti la probabilité à posteriori qu'il soit de mauvaise qualité. La décision d'accepter un bâtiment (en le considérant comme de bonne qualité) ou de le rejeter (mauvaise qualité) dépend du contexte d'application émanant du choix de l'utilisateur, qui doit définir un seuil (sur la distance radiale) à partir duquel il considère qu'un batiment n'est plus acceptable et doit être étiqueté comme étant de mauvaise qualité. Tout en laissant le choix à l'utilisateur, nous proposons deux approches pour choisir le seuil de décision.

Dans un premier temps, on se place dans un contexte du Maire de la ville qui souhaite étudier la qualité du bâti de la ville avec l'exigence de ne pas dépasser une certaine valeur sur le risque maximale de se tromper sur la détection d'un batiment de mauvaise qualité. Voyant son budget limité, il souhaite fixer un seuil maximal sur la proportion des faux positifs et cherche à maximiser le taux de vrais positifs. Cela revient à choisir sur la courbe de ROC, une valeur de *FPR* et d'en déduire la valeur de *TPR* correspondante sur la courbe. Par exemple, le Marie peut spécifier qu'il accepte 5% de faux positifs et cherche le seuil qui maximise la proportion des vrais positifs. Cette démarche se base sur la méthode communément connue sous le nom de *Constrained Maximum Success Rate* que l'on pourrais traduire littéralement par *Taux de réussite maximal contraint*, c'est-à-dire rechercher le seuil qui maximise un score sous une contrainte à ne pas dépasser. Une fois ce seuil obtenu à travers la courbe de ROC, il est possible de l'appliquer sur les probabilités a posteriori sur les bâtiments pour enfin décider de ceux étiquetés comme étant de bonne

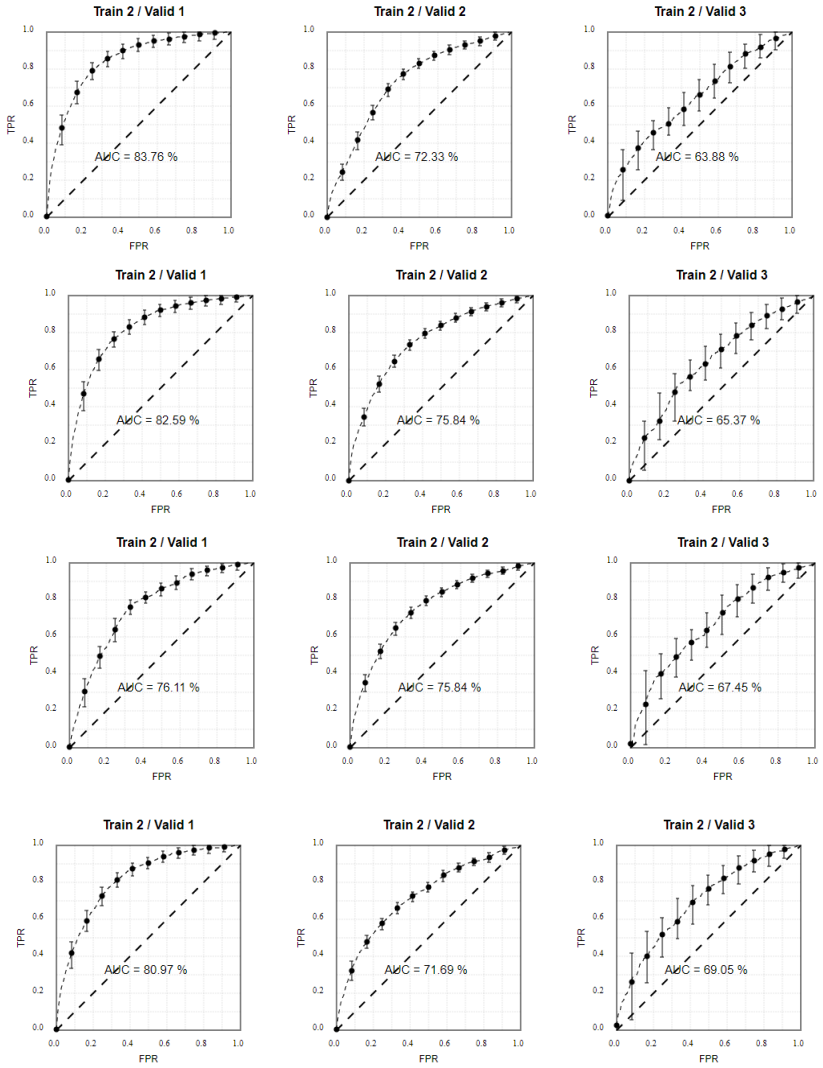


FIGURE 7.18 – Quelques exemples illustratifs de la qualité de classification pour différents seuils a priori. De haut en bas, on a les résultats de classification pour des seuils a priori respectivement à 70%, 65%, 62% et 60%. On observe bien que le seuil donnant les meilleures performances est égal à 65%.

ou de mauvaise qualité. Ainsi, le Maire pourrait corriger les bâtiments de mauvaise qualité dans la base de données tout en ayant l'assurance d'avoir économisé ses ressources et les avoir concentrées sur les bâtiments nécessitant une intervention.

Dans un second lieu, dans un contexte de disponibilité de ressources, une autre approche consiste à considérer un coût à chaque fois que l'on commet une erreur de détection que ça soit une erreur provenant de la détection d'un faux positif ou celle issue de la détection d'un faux négatif (quand on rate un vrai positif), l'objectif étant de trouver un seuil minimisant l'espérance du coût. Dans la vie réelle, ce coût peut être mesurer par le coût que l'on va dépenser pour relever et corriger un batiment malgré qu'il est réellement de bonne qualité tandis qu'un cout sur le faux négatif s'exprime par par les pertes qui peuvent advenir en considérant un bâtiment comme correct alors qu'il ne l'est

pas. Mathématiquement, on parle de pénalité, une pénalité P_{FP} quand on a un faux positif et P_{FN} quand on a un faux négatif. Nous cherchons alors le seuil de la courbe ROC qui permet de minimiser l'espérance de la pénalité. Cette démarche se base sur la méthode appelé *Minimum Mean Penalty* qui se matérialise par la minimisation de la pénalité moyenne.

Si l'on prend un cas d'application relative à la méthode de *Minimum Mean Penalty*, l'objectif n'est pas de calculer les valeurs des pénalités, mais plutôt de déterminer le rapport entre P_{FP} et P_{FN}. Ainsi, si nous supposons que le prix de la réalisation d'une carte contenant N bâtiments vaut Pr_B, le prix de mesure de la qualité par batiment s'obtient par Pr_{Bs}/N (le prix qu'on a dépensé pour relever un batiment sachant que toujours ce batiment reste toujours mal relevé) et le prix de revisite ou de la relève d'un batiment par un géomètre est égale à Pr_G.
On a le rapport : $P_{FP}/P_{FN} = Pr_G/(Pr_{Bs}/N) = N * (Pr_G/(Pr_{Bs})$ et en supposant que le P_{FP}=100 euro et que $P_{FN} = 1000euro$, le rapport donne 10%. Si F_N augmente, N augmente, Pr_G augmente et si F_N diminue, Pr_{Bs} diminue.

7.1.5.2 Démarche

Dans notre cas d'application du modèle sur les données OSM de Djibouti, nous adoptons la méthode *Minimum Mean Penalty* pour déterminer un seuil de séparation pour distinguer les bâtiments de mauvaise qualité contre ceux de bonne qualité. Pour cela, nous fixons des couts unitaires sur les faux positifs et sur les faux négatifs de manière à exprimer l'un sous forme e l'autre. Nous admettons que le coût (pénalité) des faux négatifs (C_{FN}) est trois fois plus élevé que le cout des faux positifs (C_{FP}). Bien que ce choix sur le rapport entre les coûts demeure arbitraire en terme de valeur, il exprime à quel point notre algorithme de classification est rendu plus sensible par la détection de tous les bâtiments de mauvaise qualité, et que donc ne pas les détecter nous conduirai à des couts plus importantes que le cout engendré par la détection des bâtiments de bonne qualité dans la classe positive (détection de la mauvaise qualité).

Puis une fois que nous avons fixé le rapport entre C_{FP} et C_{FN}, nous recherchons sur la courbe de ROC, le point optimal minimisant l'espérance de la pénalité. A partir de ce point, nous relevons les valeurs optimaux sur TPR et FPR donnant la performance maximale que l'on peut avoir avec notre algorithme de classification, et ceci avec une appréciation globale sur la qualité de la classification donnée par la valeur de l'AUC. Ensuite, nous retrouvons le seuil correspondant au couple TPR et FPR obtenus. Ainsi, tous les bâtiments dont leur probabilité à posteriori est supérieure à ce seuil sont jugés de mauvaise qualité.

En guise de contrôle sur les bâtiments identifiés comme étant de mauvaise qualité ainsi que la validation des résultats de la classification, nous utilisons un jeu de données sur le bâti issu d'une base de données sur Djibouti dont la précision spatiale est nettement plus meilleure que notre jeu de données OSM sur lequel a porté l'évaluation. Ce jeu de données servant de contrôle appartient à une base de données issue de la numérisation des photographies aériennes de haute résolution dans un cadre de réalisation d'un projet de coopération financé par la JICA pour le compte du Service de Topographie de l'Agence Djiboutiènne de Routes (ADR) actuellement nommé Agence Nationale de Corridors de Djibouti (ANCD). Nous comparons les bâtiments OSM qualifiés de mauvaise qualité avec leurs homologues dans la base de contrôle afin de tirer des conclusions sur les particularités

de notre modèle d'apprentissage à détecter de la qualité extrinsèque de sorte à formuler quelques considérations sur les types de bâtiments dont le modèle peut être adapté.

7.1.5.3 Résultats de l'application du modèle d'apprentissage sur Djibouti

Pour étudier la qualité du bâti à Djibouti, nous appliquons le modèle d'apprentissage sur un jeu de données composé de 15000 bâtiments extraits du bâti de la base OSM sur la ville de Djibouti. Les données OSM extraites du *94* constituent les données d'entrainement tandis que les données extraites de la base OSM de Djibouti sont les données de validation. A l'issu de l'exécution de l'algorithme de classification et en accord avec la méthode de recherche de seuil de séparation décrite dans la sous-section 7.1.5.2, nous identifions point optimal sur la courbe de ROC correspondant aux valeurs de FPR et de TPR égales respectivement à 26.5% et 74.7% dont la valeur du seuil vaut 0.4 avec une pénalité moyenne de 0.259 comme illustré sur la figure 7.19. Cela signifie qu'il faut s'attendre que l'algorithme de classification peut au mieux détecter 74.7% des bâtiments de mauvaise qualité tout en les mêlant 26.5% des bâtiments de bonne qualité.

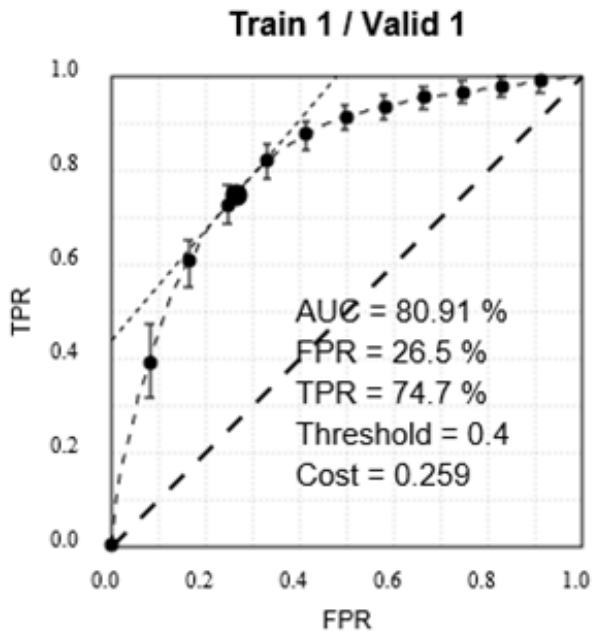


FIGURE 7.19 – Localisation du point optimal sur la courbe de ROC donnant la performance maximale du classifieur avec la minimisation de la pénalité moyenne

Afin de mieux cerner les caractéristiques des bâtiments détectés comme étant de mauvaise, nous nous intéressons à quelques cas de figure. Par exemple, nous observons un bâtiment admettant beaucoup d'irrégularités au niveau des cotés et disposant certains cotés très grands par rapport à d'autres donne ainsi un rapport du segment le plus petit sur le segment le plus long, assez faible. Cette irrégularité se manifestera aussi sur une valeur assez grande d'outlier. Ce genre de bâtiment est nettement susceptible d'être

détecté comme étant de mauvaise qualité. C'est cas de la figure 7.20. Après vérification, ce bâtiment n'existe pas en réalité, il correspond à un polygone délimitant un jardin.

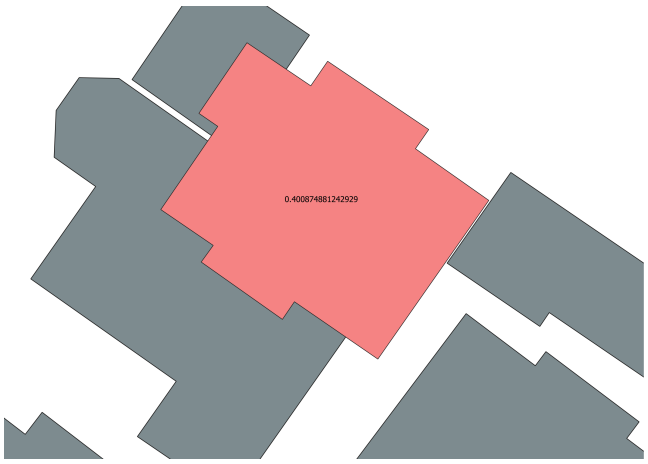


FIGURE 7.20 – Exemple de batiment (en rose) classé dans la catégorie de mauvaise qualité ayant une forme irrégulière. Les polygones en gris correspondent au bâti de la base de contrôle.

Par ailleurs, nous remarquons qu'une mauvaise qualité apparaît sur les bâtiments complexes regroupant plusieurs bâtiments. La saisie de ce type de bâtiment ne respecte pas parfois les relations topologiques avec d'autres thèmes. C'est le cas de la figure 7.21 où le bâtiment (en rouge pointillé) ne respecte pas le retrait à la route. Il chevauche l'emprise de la route.

En outre, nous remarquons que le jeu de données dans OSM, les bâtiments voisins ne se touchent. La règle de géométrie partagée n'est pas respectée contrairement aux bâtiments issus de la base de contrôle comme illustré dans la figure 7.22. Cela peut induire des erreurs de position.
De plus, les bâtiments issus du jeu d'OSM ne montrent pas un bon alignement par rapport à la route. Cela pourrait affecter également la qualité globale sur la précision de position des bâtiments comme illustré sur la figure 7.23.

Enfin, notre analyse sur les bâtiments de qualité insuffisante aboutit à la détection d'un bâtiment trop parfait constituant un morceau d'un bâtiment plus large. Sur la figure 7.24, nous observons qu'une partie du bâtiment a été saisi avec une forme en carré parfait et simplifiée.

FIGURE 7.21 – Exemple de batiment complexe (en rouge pointillé) classé dans la catégorie de mauvaise qualité. Il a été saisi en un seul polygone au lieu de deux polygones. Les polygones en gris sont issus de la base de contrôle.

FIGURE 7.22 – Exemple des bâtiments voisins du jeu d'OSM (en vert) ne disposant pas des géométries partagées. Les polygones en gris sont issus de la base de contrôle.

FIGURE 7.23 – Exemple des bâtiments ne respectant pas l'alignement par rapport à la route. Les polygones en gris sont issus de la base de contrôle.

FIGURE 7.24 – Exemple d'un bâtiment (en rouge) trop parfait constitué de la saisie d'une partie du bâtiment.

Chapitre 8

Conclusion générale

8.1 Résumé des travaux

L'information géographique a longtemps été partagée entre, producteur de données et utilisateur de données, le premier s'occupait de sa production et nécessairement de sa qualité tandis que le second se contentait de son usage. Avec l'avènement du Web 2.0, l'apparition des applications de cartographie en ligne et l'accès à des solutions mobiles, les utilisateurs contribuent à la création de la donnée géographique, et deviennent par conséquent producteurs de la donnée géographique de manière volontaire en s'organisant en communauté. Dès lors, les chercheurs étudient également la qualité de ces données dites volontaires surtout en milieu urbain où leur qualité est comparable à celle des données d'autorité. Au fil du temps, les données volontaires ont été utilisées entre autres, soit pour compléter les données d'autorité ou soit pour aider leur mise à jour grâce à une plus grande réactivité des processus collaboratifs ou encore pour servir de référence dans les zones où il n'y a pas de données d'autorité. Bien avant leur utilisation, la qualité de ces données est évalué en les comparant généralement avec des données de référence.

L'objectif de nos travaux a porté sur la mise en place un cadre d'étude permettant d'inférer la précision spatiale relative à l'aide des méthodes d'apprentissage automatique en menant uniquement une évaluation de la qualité intrinsèque basée sur les données elles-mêmes à l'état courant. Pour cela, nous avons cherché à établir un lien statistique entre des mesures extrinsèques de qualité, et des indicateurs intrinsèques de qualité pour disposer d'une estimation des mesures extrinsèques de qualité d'un jeu de données OSM pour lequel il n'y aurait pas de référence comme c'est le cas pour la République. C'est la question à laquelle nous avons cherché à apporter des éléments de réponse dans cette thèse.

Dans un premier temps, dans le chapitre 3, nous avons passé en revue les méthodes de qualification des données géographiques volontaires sous deux grandes catégories : qualité extrinsèque et qualité intrinsèque. L'évaluation de la qualité extrinsèque se fait par une comparaison d'un jeu OSM avec un jeu d'une base de référence et l'évaluation de la qualité liée aux contributeurs tandis que l'évaluation de la qualité intrinsèque s'établit à travers des indicateurs reposant sur les données elles-mêmes, actuelles, une qualité basée sur l'historique d'édition et enfin une qualité basée sur le contexte spatial.

Par la suite, en faisant l'hypothèse qu'il existe un lien statistique entre mesures extrinsèques et indicateurs intrinsèques, nous proposons une méthodologie globale de la thèse dans le chapitre 4. Globalement cette méthodologie suit une démarche qui repose sur cinq

grandes étapes successives : choix et calcul des indicateurs extrinsèques, appariement, définition et calcul d'indicateurs intrinsèques, inférence de la qualité extrinsèque à partir de la qualité intrinsèque et enfin transférabilité de l'inférence. La proposition de cette méthodologie est une des principales contributions de notre thèse.

Pour comparer deux objets surfaciques, nous identifions et choisissons d'une part deux indicateurs extrinsèques relatifs à la forme des objets surfaciques à savoir la distance radiale et la distance angulaire et deux indicateurs de position relatifs à la position des objets surfaciques à savoir la distance surfacique et la distance de Hausdorff d'autre part. A ce niveau, nous notons plusieurs apports concernant le calcul et l'implémentation des indicateurs de forme. Premièrement, nous avons recherché le facteur de sur-échantillonnage optimal des points du polygone réalisant le meilleur compromis entre vitesse de calcul et qualité de reconstruction de la signature radiale (ou angulaire).

Quant à l'étape d'appariement détaillée dans le chapitre 5, nous avons choisi une approche qui vise à être précise et robuste car l'appariement est la pierre angulaire de notre travail. Nous optons pour une méthode basée sur la théorie de croyance car cette méthode d'appariement permet de prendre en compte les cas où la connaissance peut être manquante, incertaine ou imprécise. Cela signifie que la possibilité de non-appariement est admise. Aussi, l'appariement de données basé sur la théorie de croyance autorise de fusionner des connaissances à partir de plusieurs critères d'appariement pour décider de choisir le meilleur objet homologue soutenu par tous les critères d'appariement ou d'opter pour une solution d'indécision. Ce qui convient à l'appariement des données OSM caractérisées parfois par une hétérogénéité spatiale et un manque d'attributs. A ce stade d'appariement, la valeur ajoutée de notre travail par rapport aux travaux antérieurs, apparait à deux niveaux. En premier lieu, l'appariement de données basé sur la théorie de croyance a été testé pour la première fois sur des objets surfaciques dans ce travail. En second lieu, notre apport a porté sur la définition des critères d'appariement et la détermination (par une étude empirique) des seuils adaptés à chacun de ces critères d'appariement. En fin du processus d'appariement, nous étiquetons sur les objets OSM appariés, leurs écarts avec leurs objets homologues, comme mesures de la qualité extrinsèque.

Le chapitre 6 commence par la définition d'indicateurs intrinsèques. Nous avons cherché à définir des indicateurs intrinsèques qui pourraient être corrélés avec la précision de forme et de position. Ici notre contribution a consisté à identifier dans la littérature des descripteurs capables de caractériser la qualité de saisie des objets surfaciques. Le choix d'une liste d'indicateurs non exhaustive émane d'une réflexion sur les conditions de saisie et entre dans un cadre de recherche plus général à savoir, la détection des formes du bâti par apprentissage, avec une nécessité de proposer des descripteurs adaptés. Par la suite, nous avons étendu notre réflexion sur la possibilité d'utiliser des indicateurs utilisant les caractéristiques agrégées dans un voisinage du bâtiment après avoir vérifier l'existence d'une auto-corrélation spatiale à une certaine portée des mesures de qualité. Dans ce voisinage, nous avons également défini des indicateurs capables de quantifier le degré de ressemblance d'un groupe d'objets surfaciques dans la mesure où cette ressemblance est susceptible d'être corrélée d'une manière ou d'une autre à la qualité de saisie du groupe.

Ensuite, un axe de recherche a été de chercher à établir un lien statistique significatif entre indicateurs intrinsèques et mesures extrinsèques de la qualité. C'est pourquoi nous avons proposé une méthode générique qui pourrait relier les mesures extrinsèques

de la qualité issues de l'appariement et les valeurs d'indicateurs intrinsèques calculées sur les objets appariés pour pouvoir inférer la précision spatiale connaissant uniquement les indicateurs intrinsèques. Notre travail a d'abord mis en place un modèle de régression linéaire multiple à l'échelle du bâtiment. Pour améliorer les performances et la lisibilité de la méthode d'inférence, nous avons suivi deux directions. Dans la première direction nous avons contraint le modèle de régression standard par une pénalité LASSO. Puis nous avons changé d'échelle pour l'estimation des mesures de qualité en essayant d'inférer ces mesures sur des groupes de bâtiments. Il s'agit d'un approfondissement du modèle de régression linéaire. Dans la seconde direction, nous avons proposé une méthode de classification de type forets Aléatoires visant à détecter une mauvaise qualité de saisie sur le bâti plutôt qu'une estimation numérique de cette qualité.

À l'issue du modèle d'inférence, nous avons étudié la généralisation du modèle de classification à travers un étude de transférabilité (chapitre 7). Ici, il est question de savoir à quel degré la prédiction du modèle de classification appris sur la zone de départ est performante sur d'autres zones d'études. À l'issue de notre expérimentation, nous avons remarqué que le modèle donne des résultats satisfaisants sur de nouvelles zones et même si la performance du modèle dépend des caractéristiques des données sur lesquels il est appliqué. Aussi avons cherché à déterminer si le modèle d'apprentissage s'adapte à chaque zone d'application en choisissant des nouvelles variables explicatives ou s'il existe certaines variables universelles sur lesquelles le modèle se base quelque soit la zone d'étude. Pour cela, nous avons effectué une expérimentation autour des variables explicatives afin d'étudier d'un entrainement à un autre (en changeant la zone d'étude pour l'entrainement du modèle), les variables significatives sélectionnées. Nous avons pu observer que l'algorithme de classification a choisi les mêmes variables explications pour l'ensemble des modèles d'apprentissages construits sur chacun de trois zones d'études. Ceci prouve l'universalité des variables explicatives constituant le modèle d'apprentissage.

En guise d'application, le modèle de classification retenu à l'issue de cette étude a été testé sur un jeu de données issu de la base OSM de la ville de Djibouti pour laquelle il n'existe pas des données de référence. Dans cette expérimentation, nous souhaitons répondre à la question de l'utilisation du modèle de classification en examinant le cadre applicatif du modèle pour produire une évaluation qualitative de la précision spatiale du jeu de données de Djibouti. Après l'évaluation, les bâtiments détectés comme étant de mauvaise qualité ont été confronté avec un jeu de données issu d'une autre base de contrôle afin de fournir des éléments de réponse sur les spécificités du modèle de classification sous l'hypothèse que la base de contrôle est de meilleure qualité que la base OSM à tester. A l'issue de l'étape de contrôle (comparaison des bâtiments classés de mauvaise qualité avec leurs homologues dans la base de contrôle), nous avons tiré certaines conclusions sur le comportement du modèle de classification sur le jeu de données à Djibouti.

8.2 Perspectives

À la lecture des résultats obtenus lors de ce travail lors de ce travail de thèse, nous identifions quelques axes pour poursuivre les recherches et améliorer les résultats.

Dans un premier temps, nous souhaitons améliorer notre méthode de classification. En effet, Une première étape serait de combiner des indicateurs intrinsèques calculés sur le

bâti individuel avec ceux calculées sur un groupe de bâtiments de sorte à étiqueter sur chaque bâtiment à la fois des caractéristiques individuelles, et des caractéristiques de son voisinage. Par exemple, l'écart d'alignement absolu du bâtiment avec l'alignement moyen relatif au groupe pourrait signaler des erreurs de position d'un bâtiment. Il en est de même pour des indicateurs sur la régularité et sur l'orientation d'un bâtiment par rapport à ceux de son groupe. Grâce ces indicateurs globaux, pourrions être en mesure de mieux détecter des anomalies (erreurs) de position. Parallèlement à ce modèle de classification à l'échelle individuelle du bâtiment, utilisant à la fois les indicateurs locaux et les indicateurs globaux, une alternative peut consister à entraîner le modèle d'apprentissage (de classification) sur un agrégat de bâtiments formulé autour d'un voisinage assez fin. Dans le chapitre 5, nous avons démontré qu'une agrégation des bâtiments autour d'une structure spatiale rehaussait la performance de prédiction du modèle d'apprentissage. Cela réduirait également la nécessité de refaire un tri a posteriori comme nous l'avons remarqué dans l'inférence de certains indicateurs extrinsèques dans le chapitre 6. Enfin une dernière piste d'amélioration de notre méthode de travail consisterait un processus en deux parties, avec d'abord l'application d'une méthode de classification au jeu de données. Pour la partie du jeu de données détectée comme étant de qualité insuffisante, nous nous proposons d'élaborer un modèle de régression qui estimerait une qualité extrinsèque. L'idée est d'améliorer la qualité de la régression sur les bâtiments de qualité moyenne ou mauvaise en effectuant un premier filtrage par classification.

En second lieu, une autre perspective est celle de tenter de formuler des stratégies de contrôle, de modification et/ou de ressaisie de bâtiments qualifiés comme étant de qualité insuffisante, des stratégies que l'on formalisera comme des stratégies de revisite de bâtiments. L'élaboration de telles stratégies nécessite d'étudier dans une démarche d'optimisation ou d'économie, la répartition spatiale des bâtiments de qualité insuffisante, le temps de revisite, le gain en qualité attendu, le degré probabilité que l'erreur prédite soit observée sur le terrain. Cela pourrait aboutir à un ordre de visites de bâtiments ou encore une décision sur ceux à revisiter. Cela nécessiterait entre temps de mettre en place des cartes de probabilités des réalisations prédites, des scénarios de revisite. Ces réflexions devraient aboutir à une métrique mesurant par exemple l'impact d'une revisite, et l'impact de choix d'une stratégie ou d'une autre.

Dans un cadre applicatif du modèle d'apprentissage proposé dans ce travail de thèse, nous souhaitons continuer à formaliser les démarches à entreprendre pour constituer une ébauche d'une base de référence à partir d'un jeu de données d'OSM sur Djibouti testé dans ces travaux, le but ultime étant de le comparer avec un jeu de données issu d'un autre thème (réseau routier) de l'OSM de Djibouti de sorte à utiliser le premier jeu qualifié pour qualifier le second à travers la cohérence topologique entre les deux jeux de données.

Par ailleurs d'un point de vue méthodologique, notre travail de thèse a mis en place un cadre général permettant d'inférer la qualité extrinsèque à partir d'une évaluation intrinsèque de la qualité sur des objets surfaciques. Une piste de recherche consisterait de rechercher une manière d'adopter notre approche pour évaluer la qualité d'autres types d'objets à savoir des objets ponctuels ou des objets linéaires. Cela reviendrait à définir des indicateurs extrinsèques et des indicateurs intrinsèques puis rechercher un lien statistique permettant d'estimer une qualité extrinsèque en se basant uniquement sur une évaluation intrinsèque de la qualité. Et nous sommes confrontés à des objets surfaciques de formes plus complexes, nous nous demandons de quelle manière fl faudra adapter notre approche

de sorte à mieux mesurer la qualité des lacs, des polygones d'occupation du sol ou des limites administratives.

Après avoir évalué un jeu de données OSM sur Djibouti, se pose la question de la construction d'un jeu de données de référence à partir de ces données qualifiées. Ce passage d'OSM vers une base de référence nationale implique de définir un modèle de base de données pour notre référentiel, puis de l'aligner avec celui de la base OSM. Il est nécessaire également de définir des spécifications de contenu et des métadonnées pour la base de référence qui sera établie et enfin de définir une manière de procéder de la mise à jour de cette base ainsi qu'une manière de qualifier plus tard le référentiel obtenu.

Au final, une dernière perspective consisterait à constituer un modèle d'apprentissage capable d'évaluer la qualité extrinsèque à partir d'une évaluation de la qualité intrinsèque sur un jeu de données constitué à la fois d'une intégration d'objets ponctuels, linéaires et surfaciques. Cela passe d'abord par la définition d'indicateurs intrinsèques caractérisant séparément les types d'objets pour pouvoir comprendre et maitriser individuellement l'inférence de la qualité extrinsèque à partir de ces indicateurs intrinsèques. Puis, nous serons amener à définir des indicateurs intrinsèques décrivant des caractérisations communes à la fois aux objets ponctuels, linéaires et surfaciques de sorte à fusionner ces connaissances sur des objets de types différents pour proposer finalement une manière de qualifier la qualité spatiale d'une zone de l'espace et non plus seulement d'objets individuels. Ceci demanderait d'examiner les interactions entre les différents types d'objets de sorte à mieux exploiter l'apport du contexte spatial dans un contexte de qualification des données.

Bibliographie

Abbas, I. (1994). *Base de données vectorielles et erreur cartographique : Problèmes posés par le contrôle ponctuel, une méthode alternative fondée sur la distance de Hausdorff : le contrôle linéaire*. PhD thesis.

Ai, T., Cheng, X., Liu, P., et Yang, M. (2013). A shape analysis and template matching of building features by the fourier transform method. *Computers, Environment and Urban Systems*, 41 :219–233.

Ali, A. L., Schmid, F., Al-Salman, R., et Kauppinen, T. (2014). Ambiguity and plausibility : managing classification quality in volunteered geographic information. In *Proceedings of the 22nd ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 143–152.

Alt, H. et Godau, M. (1995). Computing the fréchet distance between two polygonal curves. *International Journal of Computational Geometry & Applications*, 5(01n02) :75–91.

Antoniou, V., Morley, J., et Haklay, M. (2010). Web 2.0 geotagged photos : Assessing the spatial dimension of the phenomenon. *Geomatica*, 64(1) :99–110.

Antoniou, V. et Skopeliti, A. (2015). Measures and indicators of vgi quality : An overview. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Presented at the ISPRS Geospatial Week*, pages 345–351.

Atef, B. (2001). *Qualité géométrique des entités géographiques surfaciques Application à l'appariement et définition d'une typologie des écarts géométriques*. PhD thesis, Université de la Rochelle.

Ather, A. (2009). *A Quality Analysis of OpenStreetMap Data, M. Eng.* PhD thesis, Dissertation, Department of Civil, Environmental & Geomatic Engineering ....

Badard, T. (2000). Propagation des mises à jour dans les bases de données géographiques multi-représentations par analyse des changements géographiques. *Mémoire de thèse de doctorat en Sciences de l'Information Géographique de l'Université de Marne-la-Vallée, Marne-la-Vallée, France*.

Barrette, J., August, P., Golet, F., et al. (2000). Accuracy assessment of wetland boundary delineation using aerial photography and digital orthophotography. *Photogrammetric Engineering and Remote Sensing*, 66(4) :409–416.

Barron, C., Neis, P., et Zipf, A. (2014). A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS*, 18(6) :877–895.

Basiouka, S. et Potsiou, C. (2012). Vgi in cadastre : a greek experiment to investigate the potential of crowd sourcing techniques in cadastral mapping. *Survey Review*, 44(325) :153–161.

Basiouka, S., Potsiou, C., et Bakogiannis, E. (2015). Openstreetmap for cadastral purposes : an application using vgi for official processes in urban areas. *Survey Review*, 47(344) :333–341.

Batton-Hubert, M., Desjardin, E., et Pinet, F. (2019). *L'imperfection des données géographiques 1 : Bases théoriques*. ISTE Group.

Beeri, C., Doytsher, Y., Kanza, Y., Safra, E., et Sagiv, Y. (2005). Finding corresponding objects when integrating several geo-spatial datasets. In *Proceedings of the 13th annual ACM international workshop on Geographic information systems*, pages 87–96.

Beeri, C., Kanza, Y., Safra, E., et Sagiv, Y. (2004). Object fusion in geographic information systems. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 816–827.

Bégin, D., Devillers, R., et Roche, S. (2013). Assessing volunteered geographic information (vgi) quality based on contributors' mapping behaviours. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci*, 2013 :149–154.

Begin, D., Devillers, R., et Roche, S. (2016). The life cycle of volunteered geographic information (vgi) contributors : the openstreetmap example. In *International Conference on GIScience Short Paper Proceedings*, volume 1.

Bel Hadj Ali, A. (1997). Appariement geometrique des objets géographiques et étude des indicateurs de qualité. *Saint-Mandé (Paris), Laboratoire COGIT*.

Bel Hadj Ali, A. (2001a). Positional and shape quality of areal entities in geographic databases : quality information aggregation versus measures classification.

Bel Hadj Ali, A. (2001b). *Qualité géométrique des entités géographiques surfaciques : Application à l'appariement et définition d'une typologie des écarts géométriques*. PhD thesis, Université de Marne-la-Vallée.

Biljecki, F., Heuvelink, G. B., Ledoux, H., et Stoter, J. (2015). Propagation of positional error in 3d gis : estimation of the solar irradiation of building roofs. *International Journal of Geographical Information Science*, 29(12) :2269–2294.

Bishr, M. et Janowicz, K. (2010). Can we trust information ?-the case of volunteered geographic information. In *Towards Digital Earth Search Discover and Share Geospatial Data Workshop at Future Internet Symposium, volume*, volume 640.

Bordogna, G., Carrara, P., Criscuolo, L., Pepe, M., et Rampini, A. (2014). A linguistic decision making approach to assess the quality of volunteer geographic information for citizen science. *Information Sciences*, 258 :312–327.

Bordogna, G., Kliment, T., Frigerio, L., Brivio, P. A., Crema, A., Stroppiana, D., Boschetti, M., et Sterlacchini, S. (2016). A spatial data infrastructure integrating multisource heterogeneous geospatial data and time series : A study case in agriculture. *ISPRS International Journal of Geo-Information*, 5(5) :73.

Bostrom, H. (2007). Estimating class probabilities in random forests. In *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, pages 211–216. IEEE.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5–32.

Brovelli, M. A., Minghini, M., Molinari, M. E., Molteni, M., et al. (2016). Do open geodata actually have the quality they declare ? the case study of milan, italy.

Bruns, A. (2008). *Blogs, Wikipedia, Second Life, and beyond : From production to produsage*, volume 45. Peter Lang.

Budhathoki, N. R. et Haythornthwaite, C. (2013). Motivation for open collaboration : Crowd and community models and the case of openstreetmap. *American Behavioral Scientist*, 57(5) :548–575.

Burrough, P. A. et Frank, A. (1996). *Geographic objects with indeterminate boundaries*, volume 2. CRC Press.

Chrisman, N. (2005). Traitement de la qualité : Perspective historique. *Qualité de l'information géographique*, pages 25–35.

Ciepłuch, B., Mooney, P., et Winstanley, A. C. (2011). Building generic quality indicators for openstreetmap.

Cobb, M. A., Chung, M. J., Foley III, H., Petry, F. E., Shaw, K. B., et Miller, H. V. (1998). A rule-based approach for the conflation of attributed vector data. *GeoInformatica*, 2(1) :7–35.

Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T., et Rau, R. (2011). Osmonto-an ontology of openstreetmap tags. *State of the map Europe (SOTM-EU)*, 2011.

Coleman, D., Georgiadou, Y., et Labonte, J. (2009). Volunteered geographic information : The nature and motivation of produsers. *International journal of spatial data infrastructures research*, 4(4) :332–358.

Coleman, D. J., Sabone, B., et Nkhwanana, N. J. (2010). Volunteering geographic information to authoritative databases : Linking contributor motivations to program characteristics. *Geomatica*, 64(1) :27–39.

Costes, B. et Perret, J. (2019). A hidden markov model for matching spatial networks. *Journal of Spatial Information Science*, 2019(18) :57–89.

Craglia, M., Ostermann, F., et Spinsanti, L. (2012). Digital earth from vision to practice : making sense of citizen-generated content. *International Journal of Digital Earth*, 5(5) :398–416.

Criscuolo, L., Carrara, P., Bordogna, G., Pepe, M., Zucca, F., Seppi, R., Oggioni, A., et Rampini, A. (2016). Handling quality in crowdsourced geographic information. *European Handbook of Crowdsourced Geographic Information*, page 57.

David, B. et Fasquel, P. (1997). Qualité d'une base de données géographique : concepts et terminologie.

Davis, C. A. et Fonseca, F. T. (2007). Assessing the certainty of locations produced by an address geocoding system. *Geoinformatica*, 11(1) :103–129.

De Tré, G., Bronselaer, A., Matthé, T., Van de Weghe, N., et De Maeyer, P. (2010). Consistently handling geographical user data. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 85–94. Springer.

Devillers, R., Bédard, Y., Jeansoulin, R., et Moulin, B. (2007). Towards spatial data quality information analysis tools for experts assessing the fitness for use of spatial data. *International Journal of Geographical Information Science*, 21(3) :261–282.

Devogele, T. (1997). *Processus d'intégration et d'appariement de bases de données géographiques ; application à une base de données routières multi-échelles.* PhD thesis.

Devogele, T. (2002). A new merging process for data integration based on the discrete fréchet distance. In *Advances in spatial data handling*, pages 167–181. Springer.

Deza, E. et Deza, M. (2006). Dictionary of distances,–amsterdam. *The Netherlands : Elsevier.*

Draper, N. R. et Smith, H. (1998). Serial correlation in the residuals and the durbin–watson test. *Applied Regression Analysis*, pages 179–203.

Dretske, F. (1981). Knowledge and the flow of information.

Duféal, M., Jonchères, C., et Noucher, M. (2016). Ecce carto-des espaces de la contribution a la contribution sur l'espace-profils, pratiques et valeurs d'engagement des contributeurs d'openstreetmap (osm).

Fan, H., Yang, B., Zipf, A., et Rousell, A. (2016). A polygon-based approach for matching openstreetmap road networks with regional transit authority data. *International Journal of Geographical Information Science*, 30(4) :748–764.

Fan, H., Zipf, A., Fu, Q., et Neis, P. (2014). Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science*, 28(4) :700–719.

Fisher, P., Comber, A., et Wadsworth, R. R. (2005). *Nature de l'incertitude pour les données spatiales.*

Fisher, P. F. (1999). Models of uncertainty in spatial data. *Geographical information systems*, 1 :191–205.

Fisher, P. F. et Tate, N. J. (2006). Causes and consequences of error in digital elevation models. *Progress in physical Geography*, 30(4) :467–489.

Flanagin, A. J. et Metzger, M. J. (2000). Perceptions of internet information credibility. *Journalism & mass communication quarterly*, 77(3) :515–540.

Flanagin, A. J. et Metzger, M. J. (2008). The credibility of volunteered geographic information. *GeoJournal*, 72(3-4) :137–148.

Foody, G., See, L., Fritz, S., Mooney, P., Olteanu-Raimond, A.-M., Fonte, C. C., et Antoniou, V., editors (2017). *Mapping and the Citizen Sensor.* Ubiquity Press, London.

Franzosa, R. (1995). On the equivalence of topological relations, internat. *J. Geographical Inform. Syst*, 9 :133–152.

Friedman, J., Hastie, T., et Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York, NY, USA :.

Fritz, S., See, L., et Brovelli, M. (2017). Motivating and sustaining participation in vgi.

Galloway, A. W., Tudor, M. T., et HAEGEN, W. M. V. (2006). The reliability of citizen science : a case study of oregon white oak stand surveys. *Wildlife Society Bulletin*, 34(5) :1425–1429.

Garcia-Martí, I., Zurita-Milla, R., Swart, A., van den Wijngaard, K. C., van Vliet, A. J., Bennema, S., et Harms, M. (2017). Identifying environmental and human factors associated with tick bites using volunteered reports and frequent pattern mining. *Transactions in GIS*, 21(2) :277–299.

Genet, K. S. et Sargent, L. G. (2003). Evaluation of methods and data quality from a volunteer-based amphibian call survey. *Wildlife Society Bulletin*, pages 703–714.

Genuer, R., Poggi, J.-M., et Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern recognition letters*, 31(14) :2225–2236.

Girres, J.-F. (2012a). *Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques*. PhD thesis, Université Paris-Est.

Girres, J.-F. (2012b). Modèle d'estimation de l'imprécision des mesures géométriques de données géographiques. application aux mesures de longueur et de surface.

Girres, J.-F. et Touya, G. (2010). Quality assessment of the french openstreetmap dataset. *Transactions in GIS*, 14(4) :435–459.

Goodchild, M. F. (2007). Citizens as sensors : the world of volunteered geography. *GeoJournal*, 69(4) :211–221.

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

Graser, A., Straub, M., et 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 a ustrian reference graph. *Transactions in GIS*, 18(4) :510–526.

Gregorutti, B. (2015). *Forêts aléatoires et sélection de variables : analyse des données des enregistreurs de vol pour la sécurité aérienne*. PhD thesis, Université Pierre et Marie Curie-Paris VI.

Griffith, D. A. (1989). Spatial regression analysis on the pc : spatial statistics using minitab.

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(4) :682–703.

Haklay, M., Basiouka, S., Antoniou, V., et Ather, A. (2010). How many volunteers does it take to map an area well ? the validity of linus' law to volunteered geographic information. *The cartographic journal*, 47(4) :315–322.

Haklay, M., Singleton, A., et Parker, C. (2008). Web mapping 2.0 : The neogeography of the geoweb. *Geography Compass*, 2(6) :2011–2039.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, 29(2) :147–160.

Hangouët, J.-F. (2005). Évaluation et documentation de la qualité. *Robert Devillers Rodolphe et Jeansoulin, editeur, Qualité de l'information géographique, Information géographique et aménagement du territoire*, pages 247–272.

Hashemi, P. et Abbaspour, R. A. (2015). Assessment of logical consistency in openstreetmap based on the spatial similarity concept. In *Openstreetmap in giscience*, pages 19–36. Springer.

Hausdorff, F. (1937). Set theory, chelsea, new york, 1957. *HausdorffSet theory1957*.

Hayat, F. (2019). Geoweb. 2.0 et representation carthographique.

Heipke, C. (2010). Crowdsourcing geospatial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(6) :550–557.

Helbich, M., Amelunxen, C., Neis, P., et Zipf, A. (2012). Comparative spatial analysis of positional accuracy of openstreetmap and proprietary geodata. *Proceedings of GI_Forum*, 4 :24.

Huh, Y., Yu, K., et Heo, J. (2011). Detecting conjugate-point pairs for map alignment between two polygon datasets. *Computers, Environment and Urban Systems*, 35(3) :250–262.

Inoue, R., Ishiyama, R., et Sugiura, A. (2018). Identification of geographical segmentation of the rental apartment market in the tokyo metropolitan area (short paper). In *10th International Conference on Geographic Information Science (GIScience 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Ivanovic, S. (2018). *Quality based approach for updating geographic authoritative datasets from crowdsourced GPS traces*. PhD thesis.

Ivanovic, S. S., Olteanu-Raimond, A.-M., Mustière, S., et Devogele, T. (2019). A filtering-based approach for improving crowdsourced gnss traces in a data update context. *ISPRS International Journal of Geo-Information*, 8(9) :380.

Jackson, S. P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., et Stefanidis, A. (2013). Assessing completeness and spatial error of features in volunteered geographic information. *ISPRS International Journal of Geo-Information*, 2(2) :507–530.

Jacobs, K. T. et Mitchell, S. W. (2020). Openstreetmap quality assessment using unsupervised machine learning methods. *Transactions in GIS*, 24(5) :1280–1298.

Jakobsson, A. et Giversen, J. (2007). Guidelines for implementing the iso 19100 geographic information quality standards in national mapping and cadastral agencies. *Eurogeographics Expert Group on Quality*.

John, S., Hahmann, S., Rousell, A., Löwner, M.-O., et Zipf, A. (2017). Deriving incline values for street networks from voluntarily collected gps traces. *Cartography and Geographic Information Science*, 44(2) :152–169.

Joliveau, T. (2011). Le géoweb, un nouveau défi pour les bases de données géographiques. *LEspace geographique*, 40(2) :154–163.

Joliveau, T., Noucher, M., et Roche, S. (2013). La cartographie 2.0, vers une approche critique d'un nouveau régime cartographique. *L'Information géographique*, 77(4) :29–46.

Jolivet, L. et Olteanu-Raimond, A.-M. (2017). Crowd and community sourced data quality assessment. In *International Cartographic Conference*, pages 47–60. Springer.

Kalantari, M. et La, V. (2015). Assessing openstreetmap as an open property map. In *OpenStreetMap in GIScience*, pages 255–272. Springer.

Karimipour, F. et Azari, O. (2015). Citizens as expert sensors : One step up on the vgi ladder. In *Progress in Location-Based Services 2014*, pages 213–222. Springer.

Keßler, C., Trame, J., et Kauppinen, T. (2011). Tracking editing processes in volunteered geographic information : The case of openstreetmap. In *Identifying objects, processes and events in spatio-temporally distributed data (IOPE), workshop at conference on spatial information theory*, volume 12, pages 6–8.

Kim, J. O., Yu, K., Heo, J., et Lee, W. H. (2010). A new method for matching objects in two different geospatial datasets based on the geographic context. *Computers & Geosciences*, 36(9) :1115–1122.

Klir, G. et Yuan, B. (1995). *Fuzzy sets and fuzzy logic*, volume 4. Prentice hall New Jersey.

Koukoletsos, T., Haklay, M., et Ellul, C. (2012). Assessing data completeness of vgi through an automated matching procedure for linear data. *Transactions in GIS*, 16(4) :477–498.

Kounadi, O. (2009). Assessing the quality of openstreetmap data. *Msc geographical information science, University College of London Department of Civil, Environmental And Geomatic Engineering*, page 19.

Kuai, X., Li, L., Luo, H., Hang, S., Zhang, Z., et Liu, Y. (2016). Geospatial information categories mapping in a cross-lingual environment : A case study of "surface water" categories in chinese and american topographic maps. *ISPRS International Journal of Geo-Information*, 5(6) :90.

Lambert, N. et Zanin, C. (2012). Openstreetmap : collaborer pour faire des cartes. *Mappemonde*, 107(3).

Leszczynski, A. et Wilson, M. W. (2013). Guest editorial : Theorizing the geoweb. *GeoJournal*, 78(6) :915–919.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710.

Li, D., Zhang, J., et Wu, H. (2012). Spatial data quality and beyond. *International Journal of Geographical Information Science*, 26(12) :2277–2290.

Li, L. et Goodchild, M. (2012). Automatically and accurately matching objects in geospatial datasets. *Adv. Geo-Spat. Inf. Sci*, 10 :71–79.

Lodigiani, C. et Melchiori, M. (2016). A pagerank-based reputation model for vgi data. *Procedia Computer Science*, 98 :566–571.

Ma, D., Sandberg, M., et Jiang, B. (2015). Characterizing the heterogeneity of the openstreetmap data and community. *ISPRS International Journal of Geo-Information*, 4(2) :535–550.

Machlup, F. et Mansfield, U. (1983). *The study of information : Interdisciplinary messages.* John Wiley & Sons, Inc.

Maisonneuve, N., Stevens, M., Niessen, M. E., Hanappe, P., et Steels, L. (2009). Citizen noise pollution monitoring.

Mascret, A., Devogele, T., Le Berre, I., et Hénaff, A. (2006). Coastline matching process based on the discrete fréchet distance. In *Progress in Spatial Data Handling*, pages 383–400. Springer.

Meneroux, Y. (2019). *Méthodes d'apprentissage statistique pour la détection de la signalisation routière à partir de véhicules traceurs.* PhD thesis, Paris Est.

Mericskay, B. (2008). Étude du processus de démocratisation de la géomatique à travers l'exemple du développement du géoweb 2.0 et de ses impacts sur les processus de géocollaboration, mémoire de master. *Université Paris*, 1.

Mericskay, B. et Roche, S. (2011). Cartographie 2.0 : le grand public, producteur de contenus et de savoirs géographiques avec le web 2.0. *Cybergeo : European Journal of Geography.*

Mericskay, B. et Stéphane, R. (2010). Cartographie et sig à l'ère du web 2.0. In *Conférence internationale de Géomatique et AnalyseSpatiale (SAGEO 2010)*, pages 228–242.

Minaei, M. (2020). Evolution, density and completeness of openstreetmap road networks in developing countries : the case of iran. *Applied Geography*, 119 :102246.

Minghini, M. et Frassinelli, F. (2019). Openstreetmap history for intrinsic quality assessment : Is osm up-to-date ? *Open Geospatial Data, Software and Standards*, 4(1) :1–17.

Mohammadi, N. et Malek, M. (2015). Artificial intelligence-based solution to estimate the spatial accuracy of volunteered geographic data. *Journal of Spatial Science*, 60(1) :119–135.

Mooney, P. et Corcoran, P. (2012a). The annotation process in openstreetmap. *Transactions in GIS*, 16(4) :561–579.

Mooney, P. et Corcoran, P. (2012b). Characteristics of heavily edited objects in openstreetmap. *Future Internet*, 4(1) :285–305.

Mooney, P. et Corcoran, P. (2012c). How social is openstreetmap. In *Proceedings of the 15th association of geographic information laboratories for europe international conference on geographic information science, Avignon, France*, pages 24–27.

Mooney, S. J., Bader, M. D., Lovasi, G. S., Teitler, J. O., Koenen, K. C., Aiello, A. E., Galea, S., Goldmann, E., Sheehan, D. M., et Rundle, A. G. (2017). Mooney et al. respond to "observing neighborhood physical disorder in an age of technological innovation". *American Journal of Epidemiology.*

Mustière, S. et Devogele, T. (2008). Matching networks with different levels of detail. *GeoInformatica*, 12(4) :435–453.

National Institute of Standards and Technology (1992). Spatial data transfer standard.

Neis, P. et Zielstra, D. (2014). Recent developments and future trends in volunteered geographic information research : The case of openstreetmap. *Future Internet*, 6(1) :76–106.

Neis, P., Zielstra, D., et Zipf, A. (2012). The street network evolution of crowdsourced maps : Openstreetmap in germany 2007–2011. *Future Internet*, 4(1) :1–21.

Neis, P., Zielstra, D., et Zipf, A. (2013). Comparison of volunteered geographic information data contributions and community development for selected world regions. *Future internet*, 5(2) :282–300.

Neis, P. et Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project—the case of openstreetmap. *ISPRS International Journal of Geo-Information*, 1(2) :146–165.

Noucher, M. (2014). À bordeaux, les cartes font le pont. *M@ ppemonde*, 115.

Olteanu-Raimond, A.-M. (2008). Appariement de données spatiales par prise en compte de connaissances imprécises.

Olteanu-Raimond, A.-M., Hart, G., Foody, G. M., Touya, G., Kellenberger, T., et Demetriou, D. (2017a). The scale of vgi in map production : A perspective on european national mapping agencies. *Transactions in GIS*, 21(1) :74–90.

Olteanu-Raimond, A.-M., Laakso, M., Antoniou, V., Fonte, C. C., Fonseca, A., Grus, M., Harding, J., Kellenberger, T., Minghini, M., et Skopeliti, A. (2017b). Vgi in national mapping agencies : Experiences and recommendations. *Mapping and the Citizen Sensor*, pages 299–326.

Olteanu-Raimond, A.-M., Mustiere, S., et Ruas, A. (2015). Knowledge formalization for vector data matching using belief theory. *Journal of Spatial Information Science*, 2015(10) :21–46.

Palsky, G. (2013). Cartographie participative, cartographie indisciplinée. *L'Information géographique*, 77(4) :10–25.

Pendyala, R. M. (2002). *Development of GIS-based conflation tools for data integration and matching*. Florida Department of Transportation Tallahassee, Fla.

Pisani, F. et Piotet, D. (2008). *Comment le web change le monde : l'alchimie des multitudes*. Pearson Education France.

Plantin, J.-C. (2009). Propriétés et usages de la cartographie numérique dans l'espace urbain : le projet la montre verte. *Laboratoire Paragraphe, Mémoire de maîtrise, Paris : Université Paris*, 8.

Poplin, A., Guan, W., et Lewis, B. (2017). Online survey of heterogeneous users and their usage of the interactive mapping platform worldmap. *The Cartographic Journal*, 54(3) :214–232.

Priedhorsky, R., Chen, J., Lam, S. T. K., Panciera, K., Terveen, L., et Riedl, J. (2007). Creating, destroying, and restoring value in wikipedia. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 259–268.

Purves, R., Edwardes, A., et Wood, J. (2011). Describing place through user generated content. *First Monday*, 16(9).

Raimond, A.-M. O. et Mustière, S. (2008). Data matching–a matter of belief. In *Headway in spatial data handling*, pages 501–519. Springer.

Roche, S., Propeck-Zimmermann, E., et Mericskay, B. (2013). Geoweb and crisis management : Issues and perspectives of volunteered geographic information. *GeoJournal*, 78(1) :21–40.

Ruas, A. (1999). The role of meso level for urban generalisation. In *Workshop on Progress in Automated Map Generalisation, ICA, Ottowa*.

Ruiz-Lendínez, J. J., Ariza-López, F. J., et Ureña-Cámara, M. A. (2016). A point-based methodology for the automatic positional accuracy assessment of geospatial databases. *Survey Review*, 48(349) :269–277.

Saalfeld, A. (1988). Conflation automated map compilation. *International Journal of Geographical Information System*, 2(3) :217–228.

Safra, E., Kanza, Y., Sagiv, Y., et Doytsher, Y. (2013). Ad hoc matching of vectorial road networks. *International Journal of Geographical Information Science*, 27(1) :114–153.

Saltelli, A., Chan, K., et Scott, E. M., editors (2000). *Sensitivity analysis*. Wiley series in probability and statistics. J. Wiley & sons, New York, Chichester, Weinheim.

Samal, A., Seth, S., et Cueto 1, K. (2004). A feature-based approach to conflation of geospatial sources. *International Journal of Geographical Information Science*, 18(5) :459–489.

Sedano, E. (2016). 'sensor'ship and spatial data quality. *Urban Planning*, 1(2) :75–87.

See, L., Estima, J., Pődör, A., Arsanjani, J. J., Bayas, J.-C. L., et Vatseva, R. (2017). Sources of vgi for mapping. *Citizen Sensor*, page 13.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A., Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M., et al. (2016). Crowdsourcing, citizen science or volunteered geographic information ? the current state of crowdsourced geographic information. *ISPRS International Journal of Geo-Information*, 5(5) :55.

Seeger, C. J. (2008). The role of facilitated volunteered geographic information in the landscape planning and site design process. *GeoJournal*, 72(3-4) :199–213.

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., et Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, 31(1) :139–167.

Servigne, S., Lesage, N., et Libourel, T. (2005). Composantes qualité et métadonnées. *Robert Devillers Rodolphe et Jeansoulin, editeur, Qualité de l'information géographique, Information géographique et aménagement du territoire*, pages 213–246.

Siebritz, L.-A. (2014). *Assessing the accuracy of openstreetmap data in south africa for the purpose of integrating it with authoritative data*. PhD thesis, University of Cape Town.

Singh, K. et Xie, M. (2008). Bootstrap : a statistical method. *Unpublished manuscript, Rutgers University, USA. Retrieved from http ://www. stat. rutgers. edu/home/mxie/RCPapers/bootstrap. pdf*, pages 1–14.

Smets, P. (1998). The transferable belief model for quantified belief representation. In *Quantified Representation of Uncertainty and Imprecision*, pages 267–301. Springer.

Smets, P. et Kennes, R. (1994). The transferable belief model. artificial inteligence (66), no. 2.

Song, W., Keller, J. M., Haithcoat, T. L., et Davis, C. H. (2011). Relaxation-based point feature matching for vector map conflation. *Transactions in GIS*, 15(1) :43–60.

Stein, K., Kremer, D., et Schlieder, C. (2015). Spatial collaboration networks of openstreetmap. In *OpenStreetMap in GIScience*, pages 167–186. Springer.

Stine, R. A. (1995). Graphical interpretation of variance inflation factors. *The American Statistician*, 49(1) :53–56.

Sturrock, H. J., Woolheater, K., Bennett, A. F., Andrade-Pacheco, R., et Midekisa, A. (2018). Predicting residential structures from open source remotely enumerated data using machine learning. *PloS one*, 13(9) :e0204399.

Sui, D., Elwood, S., et Goodchild, M. (2013). Crowdsourcing geographic knowledge : Volunteered geographic information (vgi) in theory and practice.

Tiefelsdorf, M. (1998). Some practical applications of moran's i's exact conditional distribution. *Papers in Regional Science*, 77(2) :101–129.

Tong, X., Liang, D., et Jin, Y. (2014). A linear road object matching method for conflation based on optimization and logistic regression. *International Journal of Geographical Information Science*, 28(4) :824–846.

Tong, X., Shi, W., et Deng, S. (2009). A probability-based multi-measure feature matching method in map conflation. *International Journal of Remote Sensing*, 30(20) :5453–5472.

Touya, G. et Brando-Escobar, C. (2013). Detecting level-of-detail inconsistencies in volunteered geographic information data sets. *Cartographica : The International Journal for Geographic Information and Geovisualization*, 48(2) :134–143.

Truong, Q. T., De Runz, C., et Touya, G. (2019a). Analysis of collaboration networks in openstreetmap through weighted social multigraph mining. *International Journal of Geographical Information Science*, 33(8) :1651–1682.

Truong, Q. T., Touya, G., et de Runz, C. (2019b). Le vandalisme dans l'information géographique volontairedétection de l'ig volontaire vandalisée-du concept à la détection non supervisée d'anomalie. *Revue Internationale de Géomatique*, 29(1) :31–56.

Tuomi, I. (1999). Data is more than knowledge : Implications of the reversed knowledge hierarchy for knowledge management and organizational memory. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, pages 12–pp. IEEE.

Turner, A. J. (2006). What is neogeography. *Introduction to Neogeography. O'Reilly Media, Inc.*

Ubeda, T. (1997). *Contrôle de la qualité spatiale des bases de données géographiques : cohérence topologique et corrections d'erreurs.* PhD thesis, Lyon, INSA.

Unger, D. R., Hung, I.-K., Zhang, Y., Parker, J., Kulhavy, D. L., et Coble, D. W. (2013). Accuracy assessment of perimeter and area calculations using consumer-grade global positioning system (gps) units in southern forests. *Southern Journal of Applied Forestry*, 37(4) :208–215.

Upton, G., Fingleton, B., et al. (1985). *Spatial data analysis by example. Volume 1 : Point pattern and quantitative data.* John Wiley & Sons Ltd.

US Bureau of the Budget (1947). United states national map accuracy standards.

Van Exel, M., Dias, E., et Fruijtier, S. (2010). The impact of crowdsourcing on spatial data quality indicators. In *Proceedings of the GIScience 2010 Doctoral Colloquium, Zurich, Switzerland*, pages 14–17.

Vance, D. (1997). Information, knowledge and wisdom : The epistemic hierarchy andcomputer-based information systems. *AMCIS 1997 Proceedings*, 124.

Vandecasteele, A. et Devillers, R. (2013). Improving volunteered geographic data quality using semantic similarity measurements. *ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1(1) :143–148.

Vandecasteele, A. et Devillers, R. (2015). Improving volunteered geographic information quality using a tag recommender system : the case of openstreetmap. In *OpenStreetMap in GIScience*, pages 59–80. Springer.

Vapnik, V. et Vapnik, V. (1998). Statistical learning theory wiley. *New York*, 1 :624.

Vapnik, V. N. (1995). The nature of statistical learning. *Theory.*

Viana, C. M., Encalada, L., et Rocha, J. (2019). The value of openstreetmap historical contributions as a source of sampling data for multi-temporal land use/cover maps. *ISPRS International Journal of Geo-Information*, 8(3) :116.

Viglino, J. (2009). Managing partners feedbacks through the geoweb. In *2010. Proceedings of the EuroSDR Workshop 'Crowd Sourcing for Updating National Databases' held from 20th to 21st August.*

Walter, V. et Fritsch, D. (1999). Matching spatial data sets : a statistical approach. *International Journal of geographical information science*, 13(5) :445–473.

Wenjing, T., Yanling, H., Yuxin, Z., et Ning, L. (2008). Research on areal feature matching algorithm based on spatial similarity. In *2008 Chinese control and decision conference*, pages 3326–3330. IEEE.

Wu, Z. et Palmer, M. (1994). Verbs semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics*, ACL '94, page 133–138, USA. Association for Computational Linguistics.

Xavier, E. M., Ariza-López, F. J., et Urena-Camara, M. A. (2016). A survey of measures and methods for matching geospatial vector datasets. *ACM Computing Surveys (CSUR)*, 49(2) :1–34.

Xu, Y., Chen, Z., Xie, Z., et Wu, L. (2017). Quality assessment of building footprint data using a deep autoencoder network. *International Journal of Geographical Information Science*, 31(10) :1929–1951.

Yan, Y., Feng, C.-C., Huang, W., Fan, H., Wang, Y.-C., et Zipf, A. (2020). Volunteered geographic information research in the first decade : a narrative review of selected journal articles in giscience. *International Journal of Geographical Information Science*, 34(9) :1765–1791.

Yang, S., Shen, J., Konečnỳ, M., Wang, Y., Štampach, R., et al. (2018). Study on the spatial heterogeneity of the poi quality in openstreetmap. In *Proceedings of the 7th International Conference on Cartography and GIS, Sozopol, Bulgaria*, pages 18–23.

Yang, Y. (2005). Can the strengths of aic and bic be shared ? a conflict between model indentification and regression estimation. *Biometrika*, 92(4) :937–950.

Zaman, A. (2000). Inconsistency of the breusch-pagan test. *Journal of Economic and Social Research*, 2(1) :1–11.

Zhang, J. et Goodchild, M. F. (2002). *Uncertainty in geographical information*. CRC press.

Zhang, M. et Meng, L. (2007). An iterative road-matching approach for the integration of postal data. *Computers, Environment and Urban Systems*, 31(5) :597–615.

Zhang, M., Shi, W., et Meng, L. (2005). A generic matching algorithm for line networks of different resolutions. In *Workshop of ICA commission on generalization and multiple representation computering faculty of a Coruña University-Campus de Elviña, Spain*, volume 9, pages 101–110. Citeseer.

Zhang, X., Ai, T., Stoter, J., et Zhao, X. (2014). Data matching of building polygons at multiple map scales improved by contextual information and relaxation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 92 :147–163.