UNIVERSITÉ DE TOULOUSE

MASTER 2 GÉOMATIQUE

« **Sc**I**e**nces **G**éomatiques en environne**M**ent et **A**ménagement » (SIGMA)
http://sigma.univ-toulouse.fr

RAPPORT DE STAGE

IDENTIFICATION DES SURFACES ARTIFICIALISÉES DANS LA RÉGION PACA

AIT AHTMAN SALAH-EDDINE

INRA PACA

Maître de stage : GENIAUX Ghislain

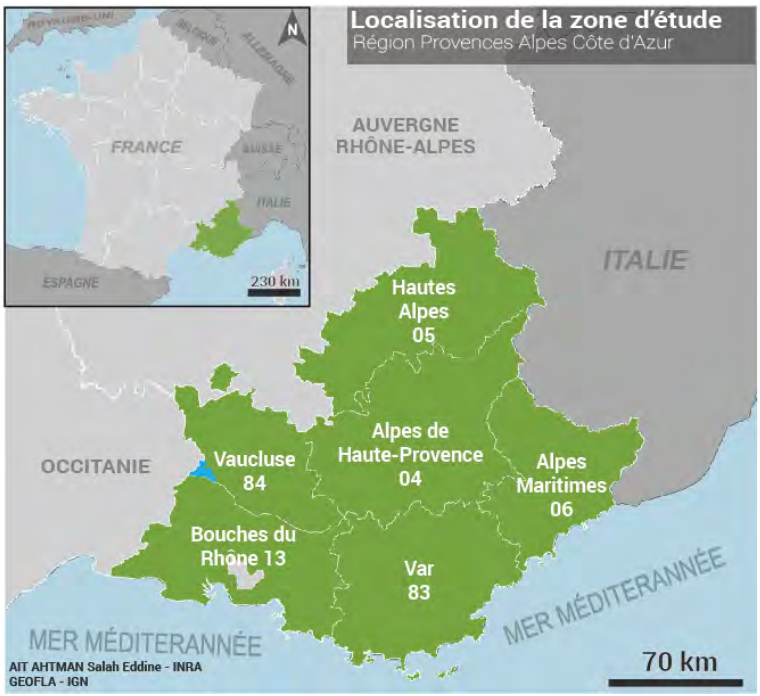
Tuteur-enseignant : FAUVEL Mathieu (ENSAT)

Septembre 2017

RÉSUMÉ

Mots-clés : Surfaces artificialisées, URBANSIMUL, OpenStreetMap, télédétection.

La croissance démographique et l'insuffisance au niveau des logements dans la région PACA, ont poussé les autorités locales à développer des stratégies visant une meilleure gestion de l'espace. L'une de ces stratégies se formalise dans la mise en place du projet URBANSIMUL. Ce dernier vise la création d'un outil d'observation et d'analyse de la disponibilité foncière à travers un ensemble de techniques statistiques et géomatiques appliquées sur des données rapatriées de diverses sources. La pertinence des résultats obtenus par cet outil dépend de la complétude et la précision des informations utilisées, c'est à ce stade qu'intervient ce stage. Il a eu pour but de compléter l'ensemble des contraintes physiques d'URBANSIMUL par les surfaces artificialisées en utilisant d'autres sources d'informations comme la base de données OpenStreetMap, ainsi que d'autres techniques d'acquisition de l'information géographique comme la télédétection. Pour arriver à cette fin, une analyse de la qualité des données OSM a été réalisée via une série de scripts python produisant des indicateurs qui décrivent l'exhaustivité et la précision géométrique de cette source par rapport à la BD TOPO de l'IGN. Cette analyse a montré que la qualité d'OpenStreetMap est acceptable puisqu'elle peut être plus exhaustive que la donnée IGN dans certains cas (zones d'activité économique), avec une différence d'exhaustivité de 43 mètres en faveur de la BD TOPO, un recouvrement supérieur à 70% et un écart géométrique de 30 mètres. Ensuite, une chaîne de traitement a été développée pour l'identification des surfaces artificialisées par télédétection à partir des images Pléiades et par l'utilisation de la méthode COO [1] et une série d'indices spectraux en excluant le bâti déjà intégré dans la base de données URBANSIMUL. L'extraction des surfaces artificialisées s'est pas sur le résultat de la classification par le modèle Random Forest (indice de Kappa : 99%), qui s'est montré plus performant que le modèle SVM (indice de Kappa : 83%) et la méthode de classification par règles de décisions sur une zone de 16Km².

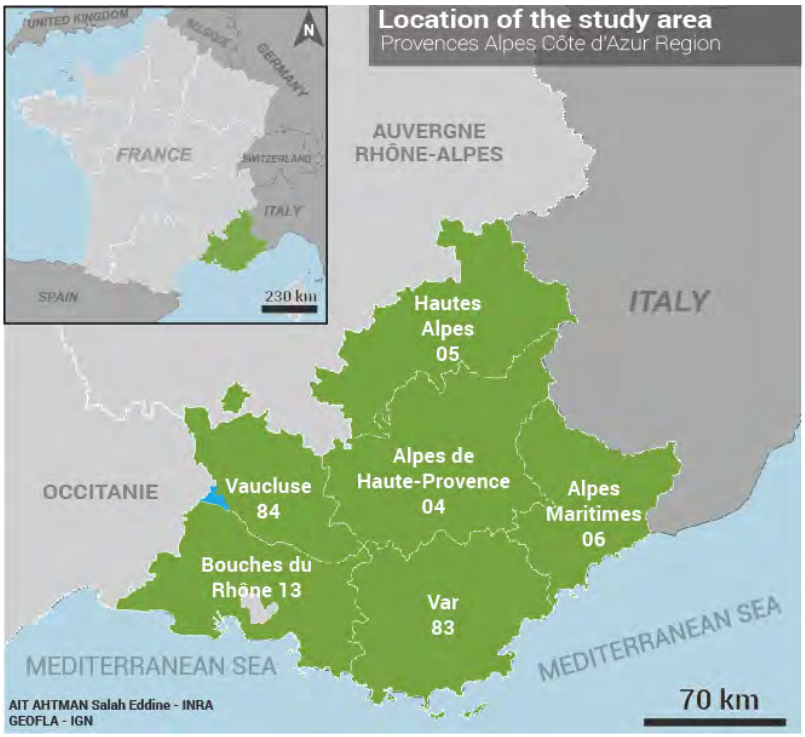


[1] COO : Classification Orientée Objet

ABSTRACT

Keywords: Artificialized areas, URBANSIMUL, OpenStreetMap, remote sensing.

The population growth and the insufficiency at the level of residences in region PACA, pushed the local authorities to develop strategies aiming a better space management. One of these strategies is formalized in the initiation of project URBANSIMUL. This last aims the creation of a tool dedicated to the observation and analysis of the land availability through statistical and geomatics techniques applied to repatriated data from various sources. The relevance of the results got by this tool depends on the completeness and the precision of information used, this is where this internship comes in. The purpose of this work was to supplement the physical constraints used in URBANSIMUL simulations by the artificialized areas extracted from other information sources like the OpenStreetMap database, as well as other techniques of acquisition of geographical information like the remote sensing. To achieve this purpose, an analysis of the quality of data OSM was carried out via a series of Python scripts producing the indicators describing the exhaustiveness and the geometrical precision of this source compared to the conventional data of IGN. This analysis showed that the quality of OpenStreetMap is acceptable since it can be more exhaustive than IGN's data, with a difference in exhaustiveness of 43 meters in favour of the BD TOPO database in some cases (Areas of economic activities), with a covering higher than 70% and one geometrical variation of 30 meters. Then, a data processing sequence was developed for the identification of the surfaces artificialized from the images Pleiades and by using the method OOC [2]based on a series of spectral indices with excluding the frame already integrated in database URBANSIMUL. The application of the method on a zone of 16km² showed that the model Random Forest (Index of Kappa: 99%) is more powerful than the model Support Vector Machine and the rule based classification.



[2] OCC : Object Oriented Classification

REMERCIEMENTS

Avant d'entamer mon rapport de projet de fin d'études, je tiens à remercier dans un premier temps, toute l'équipe pédagogique de la formation SIGMA de l'Université Jean Jaurès et l'École Nationale Supérieure Agronomique, pour avoir assuré la partie théorique de celle-ci.

Je remercie également FAUVEL Mathieu mon enseignant-tuteur pour l'aide et les conseils concernant les missions évoquées dans ce rapport, qu'il m'a apportés lors des différents suivis.

Je tiens à remercier tout particulièrement et à témoigner toute ma reconnaissance aux personnes suivantes, pour l'expérience enrichissante et pleine d'intérêt qu'elles m'ont fait vivre durant ces six mois au sein de l'INRA :

GENIAUX Ghislain, chargé de recherche en Économie, mon tuteur, pour m'avoir intégré rapidement au sein de l'unité et m'avoir accordé toute sa confiance ; pour le temps qu'il m'a consacré tout au long de cette période, sachant répondre à toutes mes interrogations, sans oublier sa participation au cheminement de ce travail, et pour m'avoir permis le prolongement de cette expérience dans le cadre d'un contrat à durée déterminée.

LEROUX Bertrand, chef de projet URBANSIMUL au CEREMA Méditerranée pour ses conseils et ses recommandations durant toute la période du stage.

Sans oublier, les autres membres de l'équipe URBANSIMUL et les autres membres de l'unité Ecodéveloppement qui m'ont fait découvrir une nouvelle passion " La pétanque".

Enfin, je remercie ma famille et mes amis pour le soutien qu'ils m'ont offert durant cette longue quête.

SOMMAIRE

INTRODUCTION

L'augmentation continue de la population dans les territoires urbains et périurbains de la région Provence-Alpes-Côte d'Azur (PACA) a obligé les autorités locales à poser des questions autour de ce phénomène qui ne cesse de s'accroître jour après jour, car selon l'INSEE[3], la région PACA connaîtra une croissance de 15% entre 2007 et 2040. Face à ce constat, la région mobilise toutes les ressources nécessaires pour une meilleure connaissance de l'occupation des sols et les dynamiques territoriales dans l'optique de canaliser cette évolution démographique.

L'un des travaux menés dans ce sens est le projet URBANSIMUL qui est en codéveloppement entre l'INRA représentée par son unité de recherche Écodéveloppement et la direction territoriale Méditerranée du CEREMA[4]. Cette collaboration a donné naissance à un outil d'aide à la décision permettant d'évaluer l'offre foncière sur l'ensemble de la région PACA en proposant à l'utilisateur un repérage automatique des parcelles disponibles pour l'urbanisation, ainsi qu'une panoplie de rapport statistique de prévision sur les dynamiques foncières et tout cela est grâce à un système modulaire en ligne. Ce dernier se base sur une multitude d'informations géographiques issues de plusieurs sources, la pertinence de ces informations joue un rôle majeur dans la définition de la précision des évaluations générées par l'outil URBANSIMUL, c'est dans l'optique d'amélioration de la qualité des résultats obtenus par l'outil que ce stage s'est effectué.

L'objectif principal de ce travail était l'élaboration des méthodes automatisées pour l'identification des surfaces artificialisées dans la région PACA, elles sont considérées comme des zones non éligibles à la construction. Pour les détecter, deux pistes ont été utilisées, la première concerne les données libres OpenStreetMap, car elles se montrent d'une grande importance puisqu'elles détiennent un nombre important d'entités géographiques et la deuxième, s'oriente vers la génération de l'information géographique à travers des techniques de télédétection. Ce rapport expose dans un premier temps les informations qui décrivent l'entité d'accueil de ce stage. Dans un second temps, il aborde les détails concernant les façons dont les deux pistes de détection des surfaces artificialisées ont été explorées à savoir les choix méthodologiques adoptés et les résultats obtenus.

[3] INSEE : Institut National de Statistiques Et d'Études Économiques
[4] CEREMA : Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement

I. PRÉSENTATION DU CONTEXTE DU STAGE

1. Présentation de la structure d'accueil

Fondé en 1946, l'institut national de recherche agronomique (INRA) est l'un des établissements publics placés sous la double tutelle du ministère de la Recherche et du ministère de l'Agriculture. À l'heure actuelle, cet institut est considéré comme l'un des piliers de la recherche au niveau européen et même mondial grâce au nombre de publications scientifiques produites par cet organisme dans le domaine environnemental.
Les principales missions de l'INRA[5] se centrent autour de l'environnement et spécifiquement l'agriculture en assurant :

- la production et la diffusion des connaissances scientifiques.
- La contribution à l'innovation par le biais des partenariats et le transfert des compétences.
- La formation à la recherche en s'appuyant sur la recherche
- la mise en place des stratégies de recherche qui cadre le niveau national et européen.
- La contribution au dialogue entre la science et la société.

Cette structure compte environ 8165 agents titulaires dont 1815 sont des chercheurs qui se répartissent sur 250 unités de recherche et 48 unités expérimentales sur l'ensemble des 17 centres régionaux de recherche. Parmi les centres régionaux de l'INRA, on trouve celui de la région PACA, il est le fruit d'une fusion entre le centre de Sophia-Antipolis et les centres d'Avignon Saint-Maurice et Saint-Paul. Il prend la 4e position au niveau national avec un budget qui dépasse les 50 millions d'euros, il compte aussi 26 unités de recherche, localisées sur 10 sites entre Avignon et Sophia-Antipolis, et huit autres sites : Aix-en-Provence, Gotheron, Le Cap d'Antibes, Nice, Les Vignères, Manduel, l'Amarine et Marseille.

Figure 1 : Schéma de la dépendance hiérarchique de la structure d'accueil

L'unité d'accueil de ce stage est l'unité Ecodéveloppement qui fait partie du centre de recherche Saint-Paul à Avignon et le département SAD[6]. Cette unité centre ses travaux autour des sciences sociales (l'anthropologie, l'économie et la sociologie), des sciences biotechniques et des sciences de la nature par l'intermédiaire de 19 agents permanents qui s'occupent principalement aux questions liées à l'écologisation durable dans les milieux agricoles et ruraux.
En outre, l'unité Ecodéveloppement s'est intéressé également au cours de ces dernières années à une thématique particulière qui est l'analyse du foncier sur l'ensemble de la région PACA à travers la création d'un outil permettant de simuler l'offre foncière nommée URBANSIMUL.

[5] www.inra.fr www.paca.fr
[6] SAD : Sciences pour l'Action et le Développement.

2. Présentation du projet URBANSIMUL

URBANSIMUL est un outil de prospection foncière développé par l'INRA et le CEREMA dans le cadre d'un partenariat avec le conseil régional de PACA, la DREAL, l'EPF et le CRIGE de PACA. Le projet donnant naissance à cet outil a été initié en 2004 dans le but de réaliser une identification qualitative des espaces libres éligibles à la construction avec un nombre limité d'informations géographiques. Le succès de cette première phase de ce projet a poussé le conseil régional de PACA et la DREAL à renouveler sa confiance en ce projet en appliquant les méthodes automatisées d'URBANSIMUL pour la détection des gisements sur trois SCoT du Vaucluse entre 2007 et 2009. À partir de 2009 la convention de ce projet a connu une évolution remarquable en matière de disponibilité de données grâce à un nouveau partenariat avec le CRIGE de la région et le CETE Méditerranée. L'évolution ne s'est pas arrêtée à ce stade-là, car le projet bénéficie jusqu'à présent et depuis la fin 2015 d'un financement dans le cadre du Plan Etat-Région 2015-2020 visant le développement de nouveaux outils d'aide à la décision et l'observation du foncier régional.
Le lancement de l'outil URBANSIMUL est prévu pour le mois de septembre 2017, il est destiné aux collectivités territoriales et les services déconcentrés de l'État qui pourront à travers les différents modules libres de cette plateforme :

- Réaliser des simulations sur l'offre foncière.
- Intégrer de nouvelles données grâce à un module d'import dans l'optique de maintenir à jour les informations stockées dans les bases de données d'URBANSIMUL.
- Produire des diagnostics touchant les processus d'urbanisation et sa régulation publique.
- Simuler les effets des différentes politiques d'urbanisation sur le territoire de la région.

À l'heure actuelle, l'outil s'appuie sur un nombre très important de données qui servent comme première base de production des contraintes physiques et réglementaires des simulations produites par cette plateforme. Ces données proviennent de plusieurs sources c'est pour cette raison que des méthodes automatisées de mise en forme ont été élaborées pour standardiser les données selon les normes du projet URBANSIMUL.

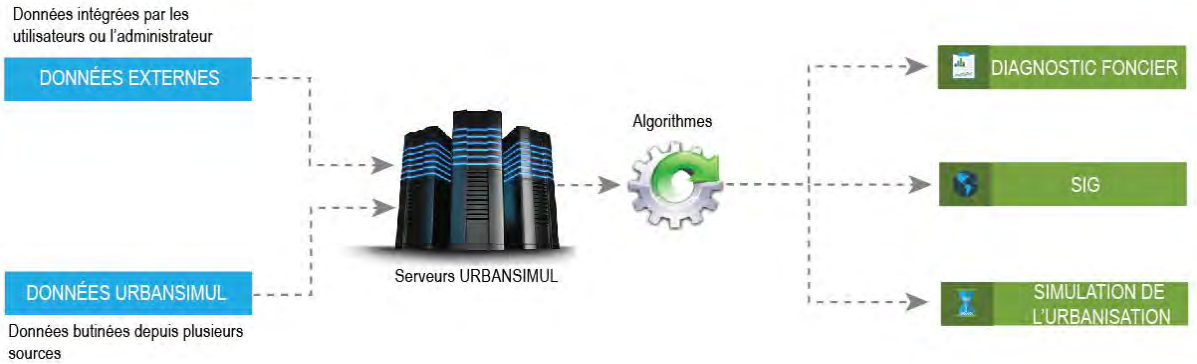


Figure 2 : Fonctionnement de l'outil URBANSIMUL

En effet, la qualité des simulations réalisées par URBANSIMUL dépend directement des données, c'est pour cette raison que la complétude des informations géographiques est nécessaire et c'est dans cette optique que le stage s'inscrit afin de compléter la base de données d'URBANSIMUL par les surfaces artificialisées.

3. Les objectifs et l'organisation du stage

a. Les objectifs

Ce stage s'inscrit dans le cadre de la phase finale du projet URBANSIMUL. Il a pour but de contribuer à l'amélioration de la qualité des rendus cartographiques de l'outil à travers la détection des surfaces artificialisées. Ces données vont permettre par la suite aux utilisateurs d'URBANSIMUL d'obtenir des résultats fiables et avec une exactitude remarquable lors de leurs simulations.
La première phase de ce stage a été dédiée à l'évaluation de l'apport de la donnée OpenStreetMap au projet par l'élaboration des méthodes d'analyse de la pertinence de l'information géographique libre afin de complémenter les données disponibles de ce projet. Ensuite, il a été nécessaire d'explorer également la piste d'acquisition de l'information à partir des images satellites qui est l'une des sources les plus importantes dans l'acquisition de la donnée géographique et cela a été matérialisé par l'élaboration d'un processus automatique d'extraction des surfaces artificialisées.
Lors des deux phases de ce stage, l'utilisation des données disponibles dans les bases de données de ce projet a été favorisée ainsi que les outils libres ou ceux dont dispose le plateau géomatique du centre INRA PACA.

Figure 3 : plan des objectifs du stage

b. L'organisation du stage

L'organisation est élément très important dans un stage, car elle permet d'avoir une vision généralisée sur le mode du fonctionnement adopté pour assurer une meilleure conduite du projet. Pour aboutir à cette fin, l'envoi d'un compte-rendu mensuel à l'encadrant universitaire Monsieur FAUVEL Mathieu contenant un diagramme de GANTT schématisant les principales tâches menées dans la période concernée. D'une autre part, des réunions ont été établies avec mon maître du stage GENIAUX Ghislain, LEROUX Bertrand qui est le chef du projet au CEREMA méditerranée et le reste de l'équipe URBANSIMUL pour mettre en avant les avancées du projet globalement et de mon stage en particulièrement.

Le temps de travail lors de ce stage a été réparti sur deux phases, la première a été consacrée à l'analyse de la pertinence et l'apport des données OSM, tandis que la deuxième phase s'est orientée plutôt vers l'extraction des surfaces artificialisées à partir des images satellites. Le travail réalisé a été réparti en neuf étapes présentées dans le diagramme de GANTT ci-dessous :

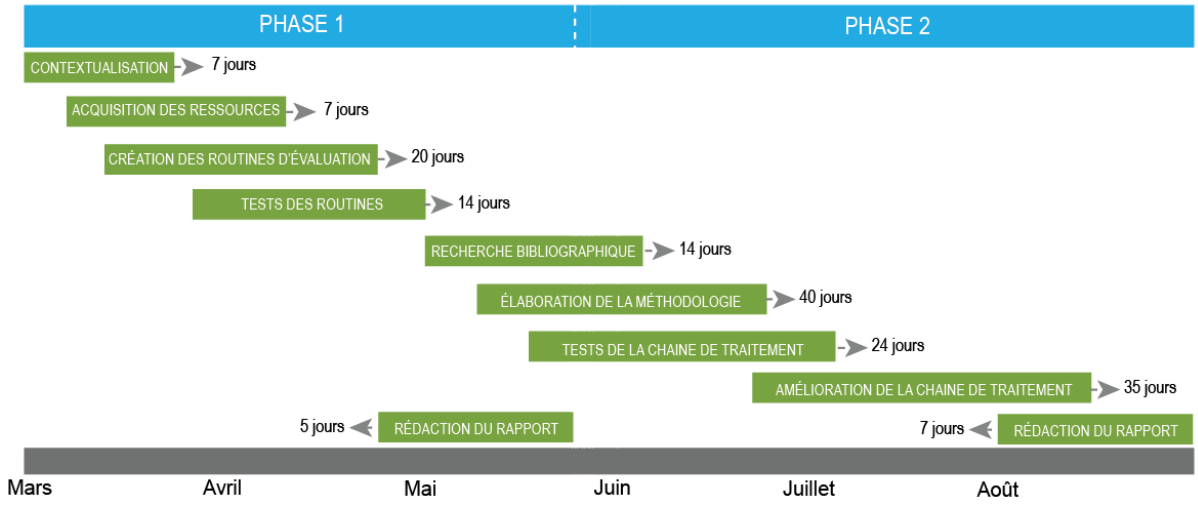


Figure 4 : Diagramme de GANTT

NB: dans le déroulement de ce stage dans certaines situations, deux étapes se chevauchent vu qu'elles ont été traitées simultanément.

PHASE 1 : l'évaluation de la donnée libre d'OpenStreetMap

Cette première s'est caractérisée par les échanges avec les chefs de projets afin de définir le plan de travail de ce stage.

Étape 1 : contextualisation

Découverte du projet URBANSIMUL, les collaborateurs dans ce projet et la prise en main de la base de données et des interfaces du projet.

Étape 2 : acquisition des ressources

Appropriation des ressources nécessaires pour l'évaluation de la qualité des données OSM.

Étape 3 : mise en place des routines d'évaluation

Création des chaînes de traitements automatisées d'évaluation de la qualité géométrique des données OpenStreetMap par comparaison avec la BD TOPO de l'IGN.

Étape 4 : rédaction du rapport

Synthétisation des résultats et les écarts observés entre les deux sources de données.

PHASE 2 : l'extraction des surfaces artificialisées par Télédétection

La deuxième phase a été consacrée pour l'élaboration d'une méthode d'extraction des surfaces artificialisées en s'appuyant sur les techniques de télédétection

Étape 1 : recherche bibliographique

Avant de se lancer dans la partie programmation de la chaîne de traitement, une recherche bibliographique a été d'une grande nécessité pour identifier les techniques précédemment élaborées pour aboutir à la même fin.

Étape 2 : élaboration d'une méthodologie

Après la recherche bibliographique, les techniques identifiées ont été intégrées dans une seule méthode dans cette partie.

Étape 3 : création de la chaîne de traitement

Dans cette étape, la méthodologie adoptée a été traduite en scripts python dans le but de l'automatiser

Étape 4 : Rédaction du rapport

Rédaction de la synthèse décrivant la méthodologie adoptée pour l'extraction des surfaces artificialisées par traitement d'images.

II. LES ZONES ARTIFICIALISÉES ET L'INFORMATION GÉOGRAPHIQUE

1. Définition des surfaces artificialisées

Pour définir les surfaces artificialisées, il nécessaire d'évoquer la définition du terme "'artificialisation" qui représente une transition touchant les espaces naturels à cause de l'expansion urbaine. L'artificialisation prend de l'ampleur au profit des espaces naturels qui sont définis comme étant des zones dotées d'un grand intérêt écologique et caractérisées par leur sensibilité et leur valeur patrimoniale et paysagère[7]. De ce fait, les surfaces artificialisées peuvent être définies comme étant l'ensemble des zones urbanisées (tissu urbain continu ou discontinu), les zones industrielles et commerciales, les réseaux de transport, les mines, carrières, décharges et chantiers, ainsi que les espaces verts artificialisés (espaces verts urbains, équipements sportifs et de loisirs), par opposition aux espaces agricoles, aux forêts ou milieux naturels, zones humides ou surfaces en eau[8].
Selon l'enquête LUCAS[9] de l'Eurostat, l'artificialisation occupe 5.2% en France dépassant la moyenne européenne de 4.1% en gagnant environ 60 000 hectares de surface chaque année sur le compte des espaces agricoles. Cette situation a poussé les autorités européennes à commencer un travail de réflexion autour de ce type d'occupations des sols en réalisant des outils de suivi et d'aide à la décision comme URBANSIMUL qui est dédiée à la région PACA. La région précédemment citée est en même temps l'une des plus urbanisées et les plus naturelles de France[10] avec 70% d'espaces naturels et 8.2% d'espaces artificialisés comme il est évident sur la figure à côté.

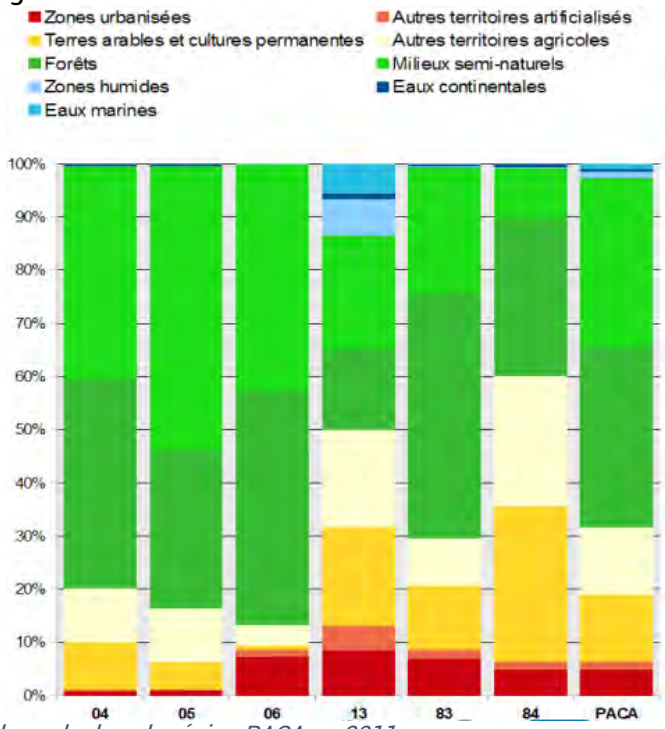


Figure 5 : L'occupation des sols dans la région PACA en 2011

[7] http://www.morbihan.fr/les-services/environnement/espaces-naturels-sensibles/les-espaces-naturels/
[8] http://www.statistiques.developpement-durable.gouv.fr
[9] http://ec.europa.eu/eurostat/web/lucas
http://www.gouvernement.fr/indicateur-artificialisation-sols
[10] http://www.paca.developpement-durable.gouv.fr

2. État de l'art

L'information géographique est devenue de nos jours la base de la prise de décision dans le monde entier, car elle offre un modèle manipulable de la réalité. Elle est utilisée depuis le début du 20e siècle pour la compréhension des différents phénomènes naturels ou anthropiques que l'environnement subit chaque jour, matérialisée au départ par des cartes sur papier puis par des systèmes virtuels dénommés les systèmes d'informations géographiques. Ce passage de l'ancienne génération à la nouvelle est dû à l'essor important connu dans les dernières années et que l'on vit toujours au jour d'aujourd'hui. À l'heure actuelle, l'acquisition de ces informations géographiques comme les surfaces artificialisées est devenue d'une simplicité remarquable grâce aux différentes plateformes et supports de distribution disponibles, le problème c'est que chacun de ces derniers présente les données d'une manière différente et avec une précision variable et ces deux éléments varient selon la provenance des données. En effet, on distingue de grandes familles d'informations géographiques, d'une part les données conventionnelles qui sont définies comme étant le fruit du travail des professionnels et des spécialistes de la création de ce type de données. D'une autre part, on trouve les informations géographiques volontaires (VGI). Elles sont éditées par des contributeurs qui ne sont pas forcément des professionnels et elles sont caractérisées par leur gratuité contrairement à la première famille qui n'est accessible gratuitement que pour une certaine catégorie[11].

La donnée libre possède plusieurs avantages comme la licence libre permettant l'exploitation et la protection des données (Baley et Touya 2014) en revanche elle présente également des inconvénients, l'un des plus impactant, c'est la précision des informations. Elle est répartie selon la littérature (Petit, Billon et Follin 2012) sur trois parties, la première est la précision attributaire définie d'après autant qu'un moyen de quantification des erreurs littérales dans la donnée. La deuxième est l'exhaustivité qui quantifie la non-représentation et la surreprésentation des objets réels et finalement, la troisième partie qui nous intéresse le plus dans le cadre ce travail avec l'exhaustivité, la précision géométrique qui a été définie comme l'estimation de l'écart entre la position réelle d'un point et sa modélisation décrite dans la donnée représentée. Cette précision géométrique a été le sujet de plusieurs études réalisées comme dans le travail de (Viry, et al. 2016) où il est précisé que l'exactitude et la complétude du jeu de données libres sont en évolution continue et dépend aussi du nombre de contributeurs actifs et leurs comportements dans la zone concernée dans le cas des données VGI d'OpenStreetMap. Cette dépendance géographique que présentent les données libres nous pousse à bien vouloir vérifier la pertinence des données de notre zone d'étude, vu qu'aucune étude n'a été menée dans ce territoire que nous étudions dans le cadre de ce stage. Les méthodes d'évaluation de la précision géométrique se basent principalement sur un ensemble de données de référence qui sont généralement des données conventionnelles comme le montre le travail de (Haklay 2010) mené sur Londres ainsi que les travaux de (Petit, Billon et Follin 2012) et (Auber, Billon et Petit 2012) menés sur la Sarthe où des méthodes de recouvrement géométrique et de mise en évidence spatiale des écarts ont été utilisées.

Les données libres ne sont pas l'unique source d'informations géographiques, il existe des processus par lesquels l'information géographique telle que les surfaces artificialisées peut être acquise, notamment la télédétection spatiale qui est définie comme l'ensemble des techniques et méthodes permettant d'étudier les phénomènes ou les objets du globe terrestre à distance à partir d'avions, de ballons ou de satellites (Kergomard, La télédétection aéro-spatiale : une introduction 1990).

L'extraction des surfaces artificialisées par le biais de la télédétection est devenue une chose atteignable grâce aux avancées technologiques et à l'ensemble de méthodes

[11] Les institutions publiques, les organismes conventionnés.

développées pour traiter les images satellites. L'un des exemples concrets présentant l'évolution de cette discipline est les images à très haute résolution spatiale qui sont de plus en plus utilisées dans la cartographie des milieux urbains (Pacifici, Chini et Emery 2009). Ce type de milieu présente sur l'imagerie THRS[12] une variation spectrale très élevée, c'est d'ailleurs pour cette raison que la méthode de classification orientée objet a vu le jour afin de traiter le contenu de l'image d'une façon différente en groupant les pixels de cette dernière en objets homogènes ou ce que l'on appelle aussi par des segments (Hamilton, et al. 2007), ce processus est dénommé la Segmentation. Plusieurs méthodes de segmentation ont été développées l'une des plus stables est la version modifiée de l'algorithme Meanshift que le CNES[13] a pu développer (Michel, Youssefi et Grizonnet 2015).

Il est nécessaire de rappeler que le point fort de la classification orientée objet, c'est qu'elle permet l'introduction des entrées comme des critères de segmentation autres que les caractéristiques spectrales de base, les indices spectraux (Jabari et Zhang 2013) ou des indices de formes (Maboudi, Amini et Hahn 2016) à titre d'exemple. Les caractéristiques citées précédemment, sont réalisées pour une classification comme dans les études de (Jabari et Zhang 2013) qui s'appuie sur la logique floue pour discriminer entre les différentes classes d'occupation des sols (Qian, et al. 2015) où plusieurs algorithmes d'apprentissage ont étés utilisés (SVM, DT …).
Les différentes méthodes d'identification des surfaces artificialisées donnent des résultats généralement satisfaisants, il reste à trouver une combinaison de traitements qui s'adapte avec les données dont on dispose dans cadre de ce projet.

3. Zone d'étude

La zone d'étude de ce travail correspond au territoire de la région française Provence-Alpes-Côte d'Azur, située en sud-est du pays elle compte 31 400km² en superficie. Elle est formée de 6 départements (Var, Vaucluse, Hautes-Alpes, Alpes-Maritimes, Bouches-du-Rhône) avec 963 communes et avec un compte d'environ 5 millions d'habitants ce qui lui donne la 7e position dans le classement des régions les plus peuplées selon l'INSEE[14] en 2016.

Figure 6 : Localisation de la région d'étude.

PACA l'une des régions qui se préoccupe beaucoup de l'aménagement du territoire à travers ces actions de lutte contre l'étalement urbain et le mitage du territoire en encourageant les recherches comme dans le cas du projet URBANSIMUL à fin de maîtriser la consommation

[12] THRS : Très Haute Résolution Spatiale ou VHR dans la littérature (Very High Resolution)
[13] CNES : Centre National d'Études Spatiales
[14] INSEE : Institut National de la Statistique et des études économiques

des espaces et préserver les milieux naturels qui représente l'une des plus importantes richesses de cette région[15].

4. Les zones artificialisées et la donnée libre

a. Présentation d'OpenStreetMap

OpenStreetMap est un projet initié en 2004 par Steve Coast un cadre universitaire dans le but de créer une base de données d'informations géographiques libre ce qui signifie que ces données sont librement utilisables et modifiables (Petit, Billon et Follin 2012). Aujourd'hui, elle est la donnée libre VGI la plus connue au monde avec une communauté qui compte des millions de contributeurs du monde entier[16].
OSM est une mine d'informations géographiques, elle est devenue en concurrence directe avec les fournisseurs des données conventionnelles parfois elle offre des informations que l'on ne peut pas retrouver dans d'autres sources ce qui fait son succès actuellement. En 2013 ce projet comptait 1 million d'utilisateurs, mais aujourd'hui le nombre d'utilisateurs à dépasser les 3 millions[17] .

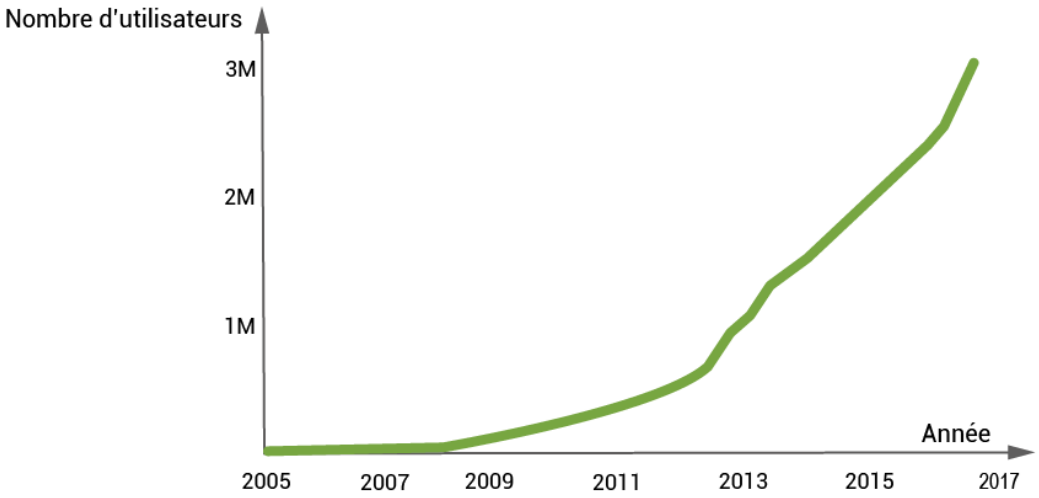


Figure 7 : Évolution des utilisateurs d'OpenStreetMap au cours des années

La structuration des données OpenStreetMap est un peu diffé-rente des autres sources de données. Chaque entité représentée est décrite par une ou plusieurs paires de types clé-valeur où la clé porte le nom du "Tag" selon la terminologie d'OpenStreetMap qui sert comme identifiant unique pour une classe d'objets par exemple :

Highway = Motorway

NB : la requête ci-dessus permet d'obtenir les entités géographiques avec l'attribut autoroute.

[15] www.regionpaca.fr
[16] www.openstreetmap.org
[17] www.openstreetmap.fr

Cette manière d'organisation des données rend ces dernières facilement exploitables à travers une panoplie d'outils dédiés à ce type de donnée dont le format de base est particulier .PBF[18] ou .OSM.
NB : Les deux formats ne sont qu'une déclinaison du langage connu sous le nom XML[19].

b. Matériel et Méthodes

i. Données utilisées

Dans cette partie, les données utilisées pour l'évaluation de la pertinence de la source OpenStreetMap au projet URBANSIMUL ont été rapatriées depuis la source GEOFABRIK qui met à disposition les données OSM avec différentes emprises géographiques à savoir des extraits continentaux, des extraits par pays et cas de la France des extraits par région. Cette flexibilité ne s'arrête pas à ces points précédemment cités, car en addition, les serveurs GEOFABRIK sont mis à jour quotidiennement et c'est la raison pour laquelle cette source a été favorisée par rapport aux autres sources d'acquisition de la donnée OSM.

Figure 8 : les formats et emprises géographiques disponibles sur GEOFABRIK

La donnée OpenStreetMap a été mise en opposition avec une source de données conventionnelles qui est dans le cas de ce travail, la BD TOPO©. Elle est une base de données vectorielle produite par l'IGN[20], elle décrit l'ensemble du territoire français avec un ensemble de données mobilisables modélisant les infrastructures linéaires de transport, le réseau de transport d'énergie, le réseau hydrographique, les bâtiments, l'occupation du sol par la végétation arborée et les limites administratives.
La construction de la BD TOPO se base principalement sur la BD CARTO et la BD ALTI de l'IGN, mais cela n'offre pas complètement une couverture globale de l'ensemble des éléments représentant la réalité. C'est donc pour cette raison que l'IGN s'appuie également sur la restitution photogrammétrique de la BD ORTHO ainsi que certaines ressources externes faibles telles que le cadastre de la DGFiP. La BD est proposée chaque année aux utilisateurs sous une nouvelle version, mais pour les missions de services publics, la recherche ou l'enseignement et l'État, elle est en libre accès avec certaines restrictions.
Au niveau de la base de données URBANSIMUL, les données disponibles pour OpenStreetMap datent d'avril 2017, pourtant pour les données BD TOPO nous avons l'historique depuis 2007 jusqu'en 2016. Dans cette partie d'analyse, c'est la version 2015 du produit IGN et la version 2017 d'OpenStreetMap qui ont été utilisées pour des raisons techniques.

[18] Protocolbuffer Binary Format
[19] Extensible Markup Language
[20] Institut national de l'information géographique et forestière

ii. Outils utilisés

Les outils utilisés pour évaluer la qualité et l'apport des données OpenStreetMap sont ceux décrits ci-dessous :

IMPOSM[21] : un outil open source d'import des données OSM avec des formats XML dans des bases de données spatiales du type PostgreSQL/PostGIS. Imposm a été développé par une compagnie allemande portant le nom OMNISCALE[22], elle met à disposition de versions de cet outil la dernière version qui est la version 3 certes elle propose de nouvelle fonctionnalité mais il n'est pas encore stable c'est pour cette raison que la version 2 a été choisi qui est stable et disponible sous les environnements Linux et Mac OS.

OGR[23] **:** une bibliothèque OGR Simple Features est une bibliothèque open source en C++ permettant la lecture et l'écriture d'une grande variété de formats de fichiers de type vecteur. Dans cette partie la fonction utilisée de cette bibliothèque est la fameuse ogr2ogr.

PYTHON[24] **:** le langage python a été utilisé principalement pour lier entre les différents outils et technologies utilisées dans l'évaluation des données OpenStreetMap.

Les paquets exploités pour aboutir à cette fin sont :

URLLIB : un ensemble de modules dédiés à l'exploitation des URL[25].

PSYCOPG 2 : paquet permettant l'accès et le requêtage d'une base de données PostgreSQL.

OS et SYS : deux paquets qui offrent la possibilité d'utiliser les fonctions natives d'un système d'exploitation.

POSTGIS[26]**:** il représente une extension de PostgreSQL permettant d'introduire la spatialisation dans une base de données ce qui veut dire que les bases de données deviennent aptes à stocker de l'information géographique.

R[27] **:** un langage de programmation a connu le jour dès 1993, dédié aux analyses statistiques particulièrement et généralement à l'analyse de la donnée scientifique. Il est structuré sous forme de paquets également et nous avons utilisé ceux décrits ci-après :

RPostgreSQL: ce paquet est l'équivalent sur R de copypg2, cité précédemment dans la partie python, il est utilisé pour assurer un accès aux données stockées sur des bases de données PostgreSQL.

---

[21] www.imposm.org
[22] www.omniscale.com
[23] www.gdal.org
[24] docs.python.org
[25] Uniform Resource Locator
[26] www.postgis.fr
[27] www.r-project.org

RANN[28] *:* paquet de détermination des proches voisins pour élément donné, la particularité de RANN, c'est qu'il inclut plusieurs distances à savoir celle de Manhattan et l'euclidienne.
Matrix: il regroupe plusieurs fonctions permettant la manipulation des matrices de données.
SELD3: un paquet interne du projet regroupant différentes fonctionnalités d'interactions spatiales, il a été développé par l'équipe URBANSIMUL.

iii. Méthodologie

La méthodologie adoptée pour l'évaluation de l'apport et la pertinence des données OSM se base sur une comparaison des surfaces artificialisées plus exactement le réseau routier d'OSM avec celui de la BD TOPO prise comme élément de référence vu qu'elle est considérée d'une bonne qualité selon l'analyse menée par le CERTU[29].Il est essentiel de rappeler que l'idée de la comparaison est issue de la littérature (Wang, et al. 2013), où la comparaison a été établie par l'utilisation de plusieurs indicateurs.

Avant d'aborder les indicateurs choisis pour cette analyse, une étape d'acquisition des données a été réalisée en élaborant un script python qui gère la récupération des données depuis la source GEOFABRIK grâce à la bibliothèque URLLIB pour les injecter par la suite dans la base de données URBANSIMUL en s'appuyant sur IMPOSM 2 et PSYCOPG 2.

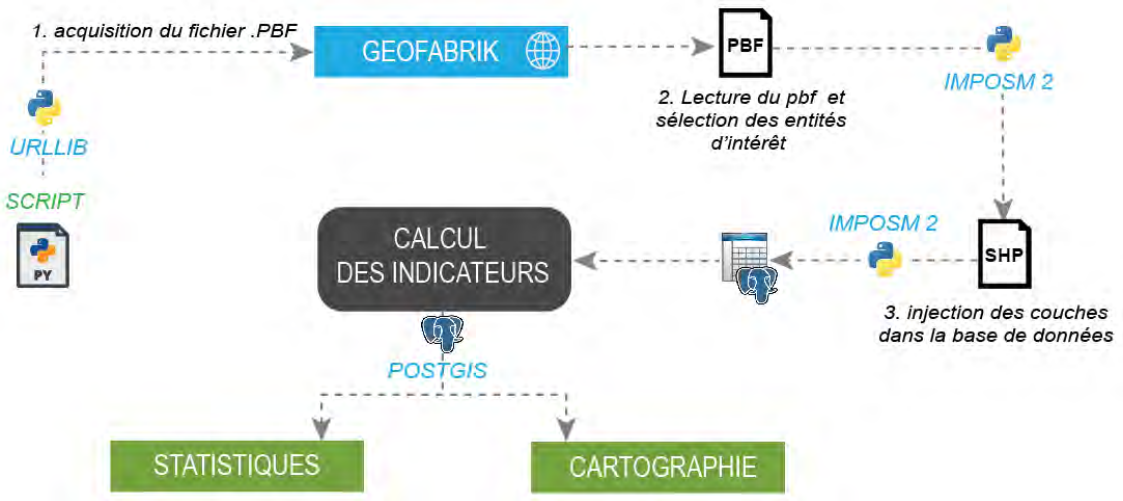


Figure 9 : Méthodologie d'évaluation des données OSM

En effet, le script python lance tout d'abord une requête d'interrogation au serveur GEOFABRIK pour demander la dernière mise à jour des données de la région PACA, ce dernier renvoie un fichier compressé au format .PBF qui sera repris par IMPOSM 2 ensuite pour le lire d'un premier temps et puis le mettre en cache afin d'assurer une disponibilité immédiate de ces données qui seront filtrées grâce un une composante d'IMPOSM 2 dénommé le fichier mapping.py. Cet élément nous a permis de ne garder que les couches utiles dont nous avons besoin parmi toutes les données disponibles.
Ensuite, IMPOSM s'occupe de l'injection de ces éléments filtrés dans la base de données du projet, mais avec quelques biais parmi lesquels on trouve le débordement de certaines

[28] Fast Nearest Neighbour Search
[29] CERTU : Centre d'Études sur les Réseaux, les Transports, l'Urbanisme et les constructions publiques

géométries de l'emprise régionale de PACA. Pour remédier à ce problème, des requêtes SQL ont été mises en place pour vérifier l'emprise des objets et la redéfinition des attributs selon les normes du projet URBANSIMUL.

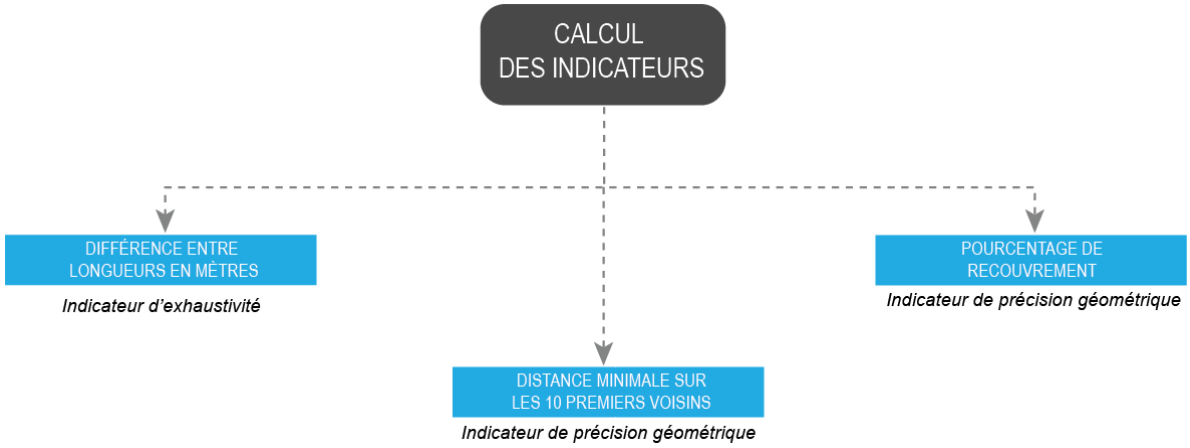


Figure 10 : Les indicateurs utilisés pour l'évaluation de la qualité des données OSM

Suite à l'acquisition des données et leur intégration dans la base de données, les indicateurs d'évaluation de la précision géométrique de la source OpenStreetMap par rapport à la BD TOPO ont été définis, premièrement nous avons utilisé la différence en mètres entre les longueurs des réseaux routiers des deux sources comme premier indicateur permettant d'obtenir une vision générale sur l'exhaustivité de la donnée OSM, cet indicateur a été utilisé dans le travail de (Petit, Billon et Follin 2012) mené sur la Sarthe.

$$Q_{exhaustivité} = L_{BDTOPO} - L_{OSM}$$

Avec :

$Q_{Exhaustivité}$: Indicateur d'exhaustivité.
L_{BDTOPO} : Longueur du tronçon routier de la BD TOPO.
L_{OSM} : Longueur du tronçon routier d'OSM.

Après la définition de l'indicateur d'exhaustivité, une évaluation de la précision géométrique a été réalisée grâce à deux indicateurs. Le premier est le pourcentage de recouvrement qui représente le rapport entre la superficie de la portion partagée entre deux entités et la superficie de l'entité de référence. Sachant que le réseau en notre possession est représenté par une géométrie linéaire pour OSM et BD TOPO, il a été nécessaire de passer par une dilatation des tronçons en fonction de type de route et de règles de dilatation qui prend en compte les largeurs réglementaires des routes afin d'obtenir une représentation surfacique des deux réseaux routiers.

$$Q_{recouvrement} = \frac{S_{INTER}}{S_{BDTOPO}}$$

Avec :

$Q_{Recouvrement}$: Indicateur de recouvrement.
S_{INTER}: La surface de la partie partagée entre la surface routière de BD TOPO et celle d'OSM.
S_{BDTOPO}: La surface routière de BD TOPO.

Enfin, le deuxième indicateur utilisé pour évaluer la précision géométrique est la distance minimale entre les objets du réseau routier BD TOPO et ceux d'OpenStreetMap. Le calcul de cette distance a été réalisé en commençant par une phase de conversion des réseaux routiers en nuages de points pour optimiser les délais de traitement. Ensuite, le résultat obtenu de cette première partie a été utilisé dans le calcul des plus proches voisins. L'idée dans cette partie est de retrouver pour chaque point du réseau routier OSM, les distances aux 10 premiers voisins du réseau IGN, le choix de 10 voisins a été fait dans le but d'analyser les écarts dans un intervalle spatial très réduit vu que les deux réseaux routiers sont très denses. Les distances calculées pour les 10 premiers voisins ont été moyennées pour obtenir la distance prise comme indicateur.
L'implémentation des deux premiers indicateurs à savoir le pourcentage de recouvrement et la différence en mètres entre les longueurs des réseaux routiers a été réalisée principalement avec des requêtes SQL lancées par un script Python qui interroge la base de données pour effectuer le calcul. En revanche, le troisième indicateur qui est la distance minimale moyennée aux 10 premiers voisins, elle a été estimée sur R grâce à la bibliothèque RPostgreSQL adoptée comme solution d'import des tables Postgres sur R et le calcul de l'indicateur a été réalisé grâce aux autres bibliothèques citées précédemment dans la partie outils utilisés. En outre, il est important de signaler que l'estimation des indicateurs d'évaluation a pris en considération la dimension spatiale en faisant le calcul sur la base d'un carroyage de 100x100m pour mettre en avant la répartition spatiale des écarts perçus sur l'ensemble de la région PACA.
Finalement, une phase de validation a été mise en place dans le but de trouver la relation qui existe entre les deux sources de données, et cela a été réalisé grâce à un croisement statistique simple entre l'indicateur d'exhaustivité et la distance aux zones commerciales qui représentent l'une des zones d'intérêt du projet URBANSIMUL.

c. *Résultats et discussions*

L'application de la chaîne d'évaluation précédemment exposée dans la partie méthodologie, nous a prouvé que les données OSM sont généralement d'une qualité qui tend vers celle de la BD TOPO. En effet, l'indicateur de l'exhaustivité a montré que globalement la source OSM est moins fournie que la BD TOPO avec un écart moyen estimé à 42m sur l'ensemble de la région, mais cette moyenne est trompeuse, car elle agrège les différents types de zonage. C'est pour cette raison qu'une évaluation a été réalisée à l'échelle des zonages qui les plus importants dans le cadre du projet URBANSIMUL à savoir les zones d'activités économiques qui ont été définies par l'intermédiaire du PLU[30].

Zonage	*Différence en m (m)*	*Recouvrement (%)*	distance min (m)
Région PACA	43	> 70	30
Z . E	-10	> 75	10

Table 1 : résumé des écarts identifiés sur l'ensemble de la région et sur les zones d'activité économique

Le résultat de l'analyse réalisée montre que les données libres d'OpenStreetMap sont d'une précision qui se rapproche beaucoup de celle de la BDTOPO. En ce qui concerne

[30] PLU : Plan Local d'Urbanisme

l'exhaustivité d'une manière globale, la BD TOPO est plus fournie que la source OSM avec une différence moyenne de 43 mètres. Ce chiffre peut paraitre énorme mais il faut mentionner que les deux réseaux routiers ne possèdent pas la même nomenclature, car il existe parfois des types des réseaux routiers qui peuvent être représentés dans une source de données mais pas dans l'autre. En recouvrement OSM couvre 70% du réseau routier de la BD TOPO et plus que 75% de ce dernier dans les zones d'activité économique et par rapport à l'écart topologique entre les représentations des deux sources, il ne dépasse pas les 30 mètres généralement.
Dans le but d'aller plus loin, une cartographie qui met en avant la répartition spatiale des écarts sur l'intégralité de la région PACA. Elle permet d'obtenir une idée sur la relation entre le positionnement des objets géographiques et la valeur des écarts.
Ci-après un exemple du rendu cartographique obtenu à l'issue de cette phase d'évaluation.

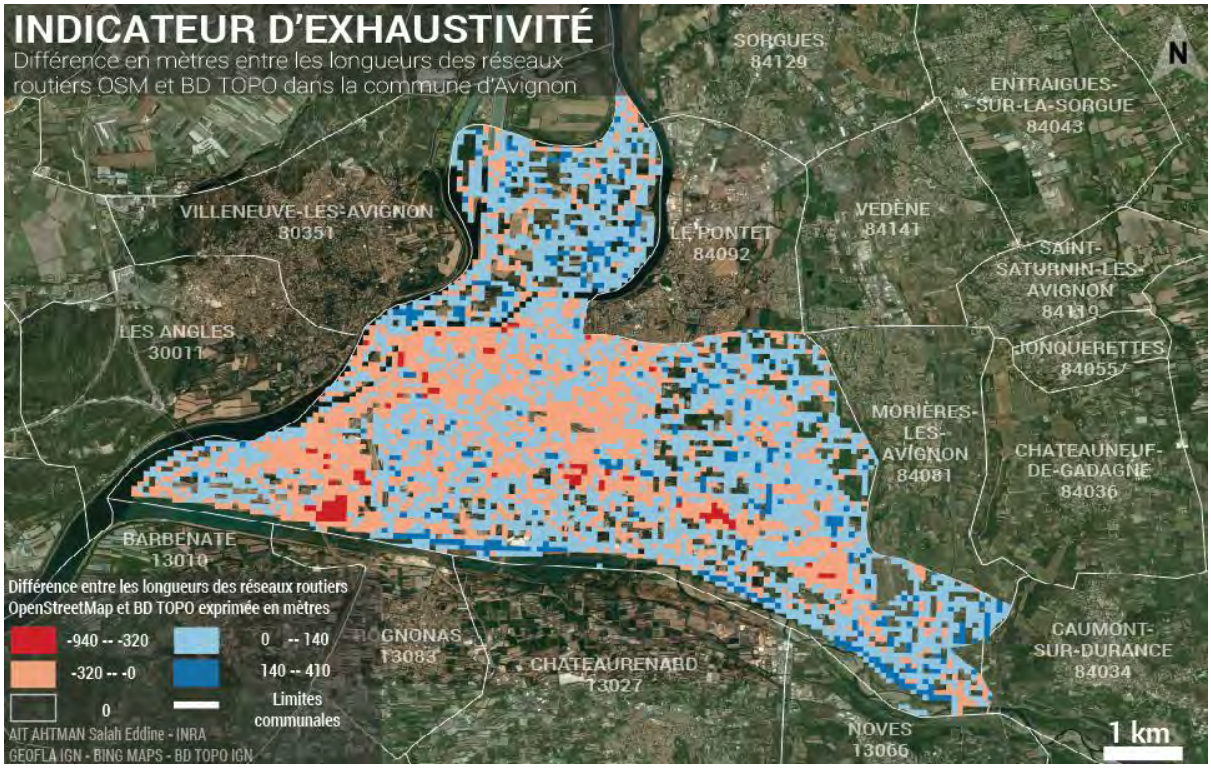


Figure 11 : Différence en mètres entre les longueurs des réseaux routiers OSM et BD TOPO sur Avignon

En effet, grâce à la cartographie des écarts, il a été constaté que OpenStreetMap offre une quantité très importante d'informations autour des zones d'activité économique comme il est évident dans la cartographie ci-dessus où nous voyons bien que les endroits où OSM est plus abondante en termes d'informations sont les zones d'activités notamment la ceinture commerciale Cristole/Mistral 7 d'Avignon. Après ce constat, nous avons posé l'hypothèse que la source OpenstreetMap peut être plus fournie que la source de référence à savoir la BD TOPO, cette dernière a été rejetée grâce au croisement réalisé entre l'indicateur d'exhaustivité et la distance aux zones d'activités commerciales.
Par rapport au deuxième indicateur qui est le pourcentage de recouvrement qui reflète la précision géométrique du réseau routier OSM, il nous a aidés a montré que le réseau OSM couvre 75% du réseau routier IGN sur toute la région PACA. Sachant que dans le cadre d'URBANSIMUL il était plus intéressant de vérifier la précision d'OSM dans les secteurs

d'activité économique nous avons pu constater que la précision est meilleure dans ce type de zonage vu qu'elle atteint 80% de recouvrement du réseau BD TOPO par celui d'OSM.
Dans l'optique de valider les résultats du dernier indicateur, un second indicateur a été élaboré comme décrit dans la partie méthodologie portant le nom de l'indicateur de distance minimale. Ce dernier a montré que l'écart entre les géométries du réseau OSM et celui de l'IGN est estimé à 13 mètres.

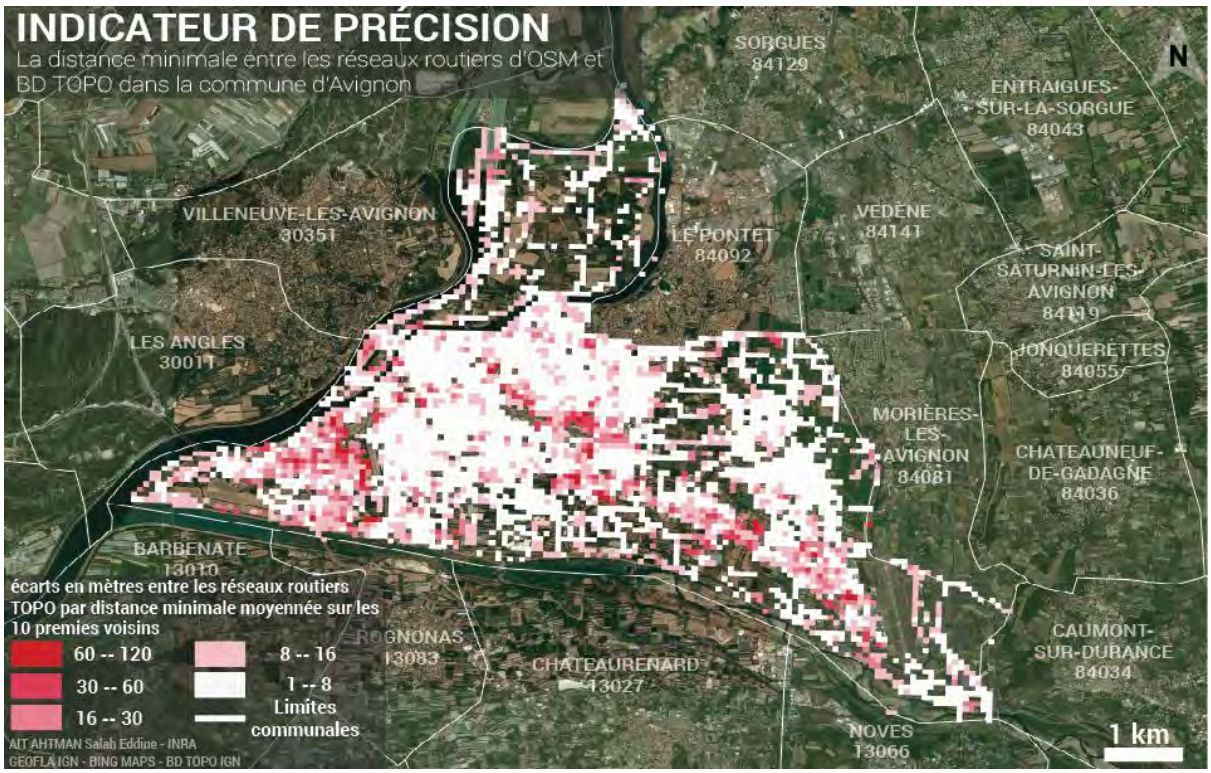


Figure 12 : la distance minimale entre les réseaux routiers OSM et BD TOPO sur la commune d'Avignon.

L'indicateur de distance minimale a été également cartographié, comme les autres indicateurs réalisés. Ci-dessous un exemple de cartographie de cet indicateur pour le cas de la commune d'Avignon.

La donnée OSM s'est avérée une source d'information très utile pour le projet URBANSIMUL, car elle permet de compléter la base de données de l'outil. Après cette analyse on a pu vérifier que la qualité de cette mine d'informations varie d'une zone à l'autre, mais globalement elle est satisfaisante dans le cas de représentation des objets géographiques inexistants dans les sources de données conventionnelles intégrées à URBANSIMUL comme le cas des parkings que l'on retrouve que dans les couches ponctuelles de la BD TOPO.
Comme il a été décrit dans le rapport technique réalisé par le CGET[31] sur les données OSM, ces derniers reste quand même assez proche en termes de précision géométrique des données conventionnelles (Zielstra et Zipf 2010).D'une autre part, certains ont montré que la donnée OSM est d'une forte hétérogénéité en matière de précision géométrique (Girres et Touya 2010), cela nous met toujours dans l'obligation de vérifier la précision de cette donnée avant son exploitation.

[31] Commissariat général à l'égalité des territoires

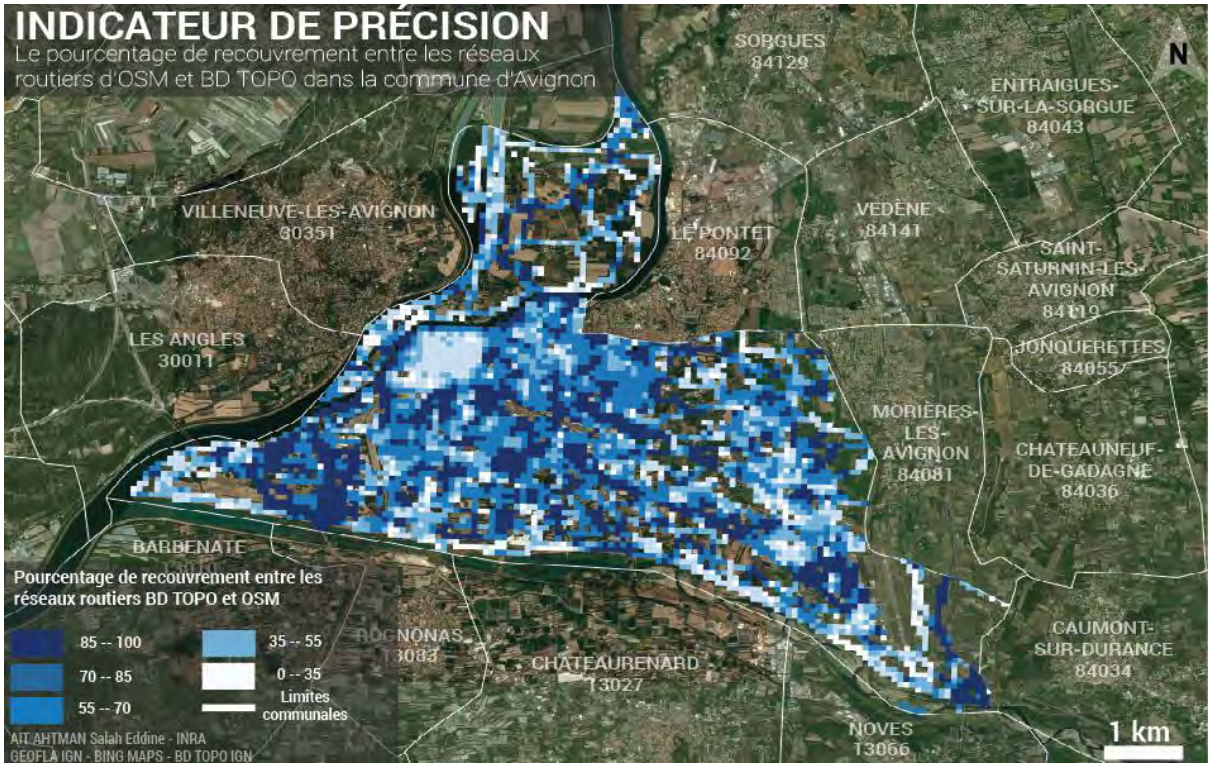


Figure 13 : le pourcentage de recouvrement entre les réseaux routiers d'OSM et BD TOPO sur Avignon

Donc il nécessaire, de voir ce que pourrait éventuellement apporter le développement des méthodes et techniques permettant la construction d'une information géographique issue de ces deux types de source, car elles peuvent être considérées comme étant deux sources complémentaires. Dans le cas d'URBANSIMUL par exemple un réseau routier détenant toutes les entités de la BD TOPO et les entités géographiques disponibles dans OpenStreetMap et qui ne sont toujours pas cartographiées sur la donnée de l'IGN. Cette approche peut être très intéressante vu que les données OSM sont caractérisées par une fréquence d'évolution importante comme nous avons pu le constater au niveau du travail de (Petit et al, 2013) où il a été évident que dans un intervalle temporel très court (1 an) le pourcentage de recouvrement a connu une évolution très significative.

5. Les zones artificialisées et la télédétection

Cette partie est consacrée à la méthode de détection des surfaces artificialisées hors bâti (bâti existant dans la base de données URBANSIMUL) élaborée par l'intermédiaire des techniques d'analyse d'images satellites notamment les images à très haute résolution spatiale.

a. Présentation de la télédétection

Selon l'Agence spatiale européenne (ESA), la télédétection est une méthode d'acquisition des informations sur les objets terrestres sans devoir être en contact direct avec ces derniers. Elle se base sur trois éléments fondamentaux :

- la plateforme : elle définit comme étant le moyen permettant à l'instrument utiliser pour les mesures de quitter la surface de la Terre. Ces porteurs de capteurs ont connu une évolution très importante grâce à l'essor technologique du 20e siècle pour passer des montgolfières et des avions aux satellites.
- l'objet : il représente l'élément observé
- Le capteur : l'appareil embarqué dans la plateforme dans le but d'acquérir des images optiques représentant différentes grandeurs physiques selon les caractéristiques de chaque capteur.

Il existe dans la télédétection deux modes de détection, le premier qui est nommé l'actif où le capteur émet une quantité d'énergie dirigée vers l'objet cible, pour mesurer ensuite l'énergie réfléchie par cet objet[32]. Dans le cadre de notre travail, nous avons utilisé l'opposée du mode actif et qui est le passif où les capteurs se basent sur le rayonnement solaire pour quantifier l'énergie émise par les objets étudiés.

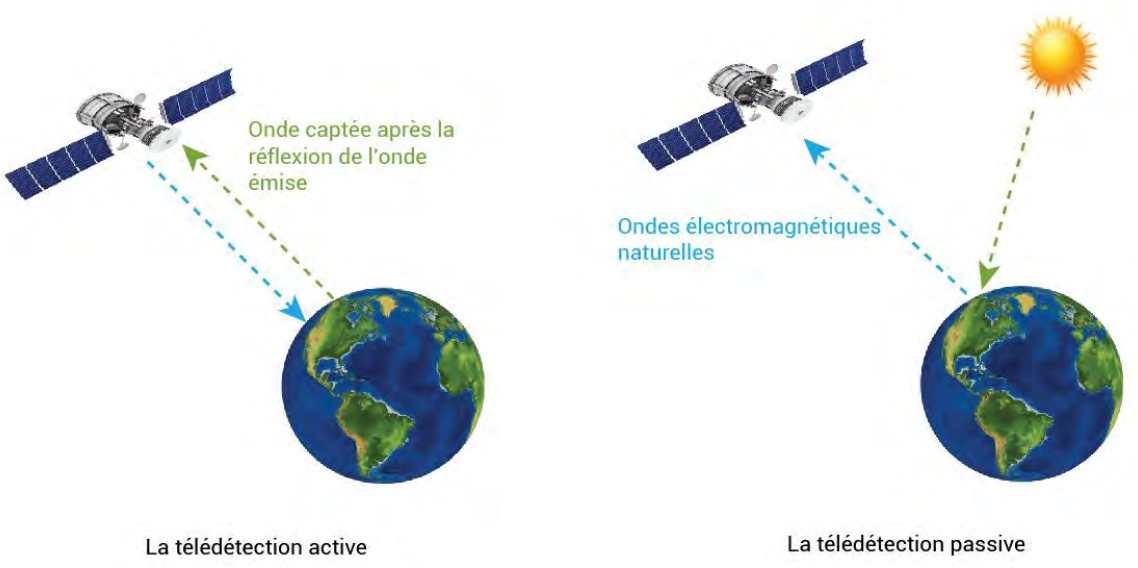


Figure 14 : Fonctionnement des modes actif et passif en Télédétection

[32] http://eoedu.belspo.be

b. Matériel et méthode

i. Données utilisées

La méthode de détection des surfaces artificialisées s'est appuyée sur les images satellites à très haute résolution spatiale qui sont les plus adaptées pour les détections touchant le milieu urbain et spécifiquement les objets à faible dimension spatiale tels que les routes, les parkings, etc. Pour obtenir ce genre de données dans le cadre du projet URBANSIMUL, il a été d'une grande nécessité de réaliser une adhésion au programme GEOSUD[33] qui est le fruit d'une collaboration entre l'AgroParisTech, l'IRSTEA[34], le CIRAD[35] et l'IRD[36] dans le domaine de l'information spatiale. Cette adhésion nous a permis d'avoir un accès libre sur les données à très haute résolution comme SPOT et Pléiades, mais le choix dans cette étude s'est porté sur les données Pléiades qui sont de meilleure résolution spatiale[37].

En effet, Pléiades est un système qui a été mis en place par le CNES sous un financement majoritaire du gouvernement français et la contribution d'autres pays tels que l'Espagne, l'Autriche, la Belgique et la Suède. Ce système a donné naissance à un couple de deux satellites identiques dédiés à l'observation de la Terre. Ils sont positionnés sur la même orbite chose qui une meilleure couverture temporelle. Le premier a été lancé en 17 décembre 2011 et le second qui le Pléiades 1B a été lancé en fin 2012.

Au sujet d'images offertes par ces satellites, elles sont disponibles en mode panchromatique et multispectral avec une différence en matière de résolution spatiale, vu que les images en panchromatique sont d'une résolution de 70cm (50cm au sol) et 2.8m pour les images multispectrales.

Bande spectrale	*Longueur d'onde*	*Résolution*
BLEU	0.43 - 0.55	2m
VERT	0.55 - 0.62	2m
ROUGE	0.59 - 0.71	2m
INFRAROUGE	0.74 - 0.94	2m
PANCHROMATIQUE	0.59 - 0.71	50cm

Table 2 : descriptif des bandes spectrales des satellites Pléiades.

La version des données utilisées dans ce projet est une version orthorectifiée où l'on ne retrouve pas des distorsions des géométries grâce à la correction réalisée par l'IGN. En plus, de la correction géométrique les bandes multispectrales ont été fusionnées avec la bande panchromatique dans le but d'améliorer la résolution spatiale sur le mode multispectral en passant de 2.8m à 0.5m.

[33] GEOinformation for SUstainable Development
[34] Institut national de Recherche en Sciences et Technologies pour l'Environnement et l'Agriculture
[35] Centre de coopération internationale en Recherche Agronomique pour le Développement
[36] Institut de Recherche pour le Développement
[37] www.ign.fr

ii. Outils utilisés

Dans la partie télédétection, les outils utilisés sont presque les mêmes que la première partie d'évaluation avec l'introduction d'un nouvel outil dédié au traitement des images et certaines nouvelles bibliothèques et tout cela est décrit ci-dessous :

OTB[38] : un outil open source développé par le CNES dédié à la manipulation des images optiques et radars grâce à une panoplie d'outils disponibles sous Monteverdi l'interface d'OrfeoToolBox, QGIS, en ligne de commande, Python et en C++.

OGR[39] **:** une bibliothèque OGR Simple Features est une bibliothèque open source en C++ permettant la manipulation des formats raster et vecteur dans cette partie les fonctions utilisées sont principalement les fonctions de lecteur, création et modification du format Raster.

PYTHON[40] **:** le langage python a été utilisé principalement pour lier entre les différents outils et technologies utilisées comme dans le cas d'évaluation des données OSM, mais cette fois-ci dans l'identification des surfaces artificialisées. Les paquets utilisés pour aboutir à cette fin sont :

SCIPY : paquet dédié aux applications scientifiques, il permet à titre d'exemple la manipulation des données au format matriciel.

NUMPY : un autre paquet dédié à la manipulation des données scientifiques.

PSYCOPG 2 : paquet permettant l'accès et le requêtage d'une base de données PostgreSQL.

OS et SYS : deux paquets qui offrent la possibilité d'utiliser les fonctions natives d'un système d'exploitation.

POSTGIS[41]**:** il représente une extension de PostgreSQL permettant d'introduire la spatialisation dans une base de données ce qui veut dire que les bases de données deviennent aptes à stocker de l'information géographique.

iii. Méthodologie

Cette partie a comme but d'exposer la méthodologie utilisée pour l'identification des surfaces artificialisées à travers des images à très haute résolution spatiale et en s'appuyant principalement sur des technologies libres telles que l'OrfeoToolBox et Postgis. L'idée dans cette partie, est de construire une chaîne de traitement qui permet l'identification des objets d'intérêt en suivant une approche orientée objet qui est la plus adaptée avec les extractions des objets urbains caractérisés par leur faible dimension spatiale.

La chaîne de traitement élaborée utilise le concept du pipeline qui est introduit dans l'OrfeoToolBox, mais avec une façon différente. Dans ce concept, plusieurs outils communiquent entre eux dans le but de compléter l'un par l'autre.

[38] OrfeoToolBox
[39] www.gdal.org
[40] docs.python.org
[41] www.postgis.fr

Ci-dessous le schéma de déroulement du processus d'identification des surfaces artificialisées:

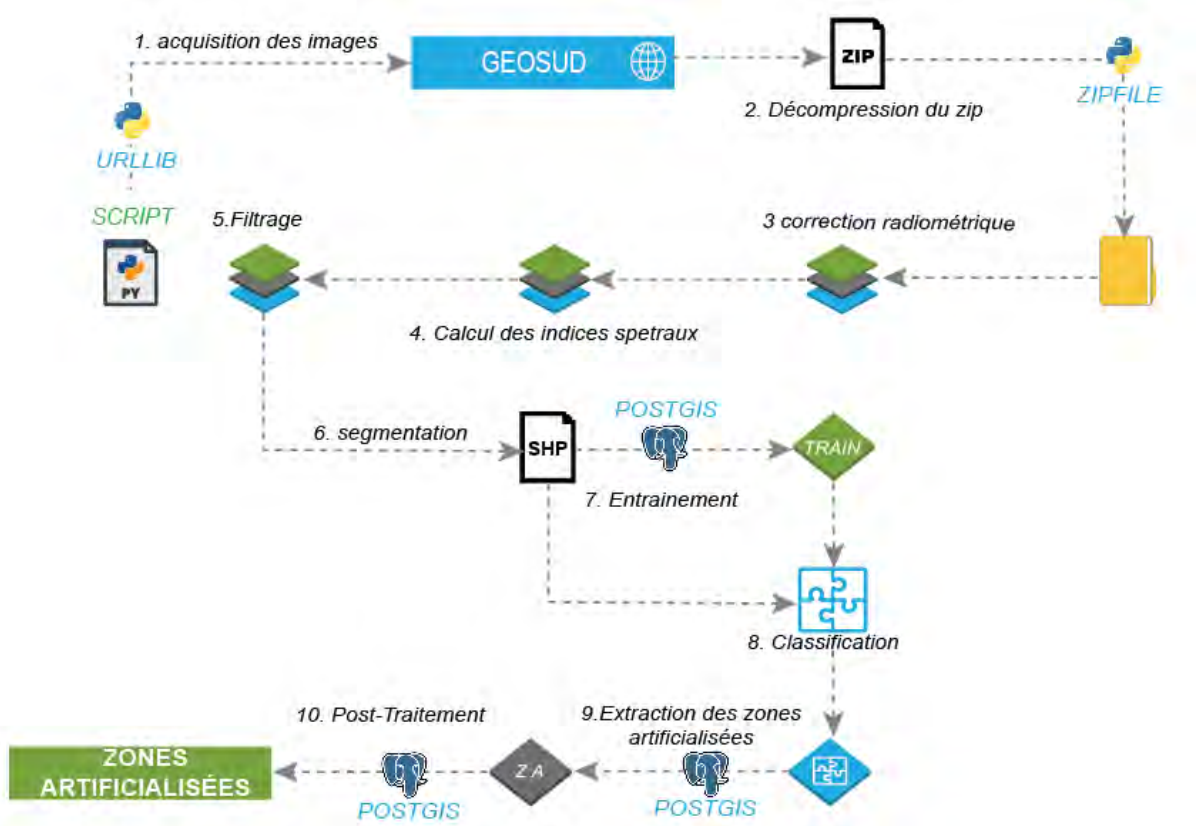


Figure 15 : Méthodologie d'extraction des zones artificialisées

L'identification des surfaces artificialisées a été réalisée suivant plusieurs étapes en commençant par une phase correction de la donnée acquise pour terminer avec une de post-traitements.

Toutes ces étapes seront décrites ci-dessus:

Correction radiométrique et atmosphérique

Les informations stockées dans les images satellitaires brutes ne peuvent pas être utilisées directement vu qu'elle présente quelques biais liés à des phénomènes naturels et à des problèmes d'étalonnage dû au capteur et à la géométrie.
En addition, la donnée brute est toujours enregistrée sous forme d'un compte numérique qui doit être converti en réflectance au sol. Pour y arriver, il est nécessaire de passer par trois étapes :

ÉTAPE 1 : elle consiste à réaliser un étalonnage absolu en fonction des coefficients de calibration du capteur ce qui permet l'élimination des effets causés par la sensibilité du capteur et également la conversion d'un compte numérique à une valeur de radiance spectrale.
La formule utilisée dans cette étape est la suivante :

$$L = \frac{DN}{GAIN} + BIAS \; en \; W/m/steradians/\mu m$$

Avec DN= le compte numérique GAIN, BIAIS: coefficients de calibrations du capteur.
ÉTAPE 2 : la seconde étape permet d'assurer le passage de la radiance, précédemment calculée au niveau du capteur, à la valeur de réflectance au sommet de l'atmosphère cette transition est réalisée grâce à un certain nombre de paramètres qui servent à ajuster la valeur de réflectance. Le calcul de cette phase se fait grâce à la formule ci-après :

$$r = \frac{\pi * L * d^2}{ESUN * cos(\theta_s)} \in [0, 1]$$

Avec:

L= radiance spectrale au niveau du capteur.
d= distance entre la terre et le soleil en unité astronomique
ESUN= constante solaire variable en fonction de la bande et du capteur
θ_s= l'angle d'éclairement zénithal

NB: il faut signaler que pour le cas des images Pléiades, il faut soustraire l'angle d'élévation du soleil à 90°.

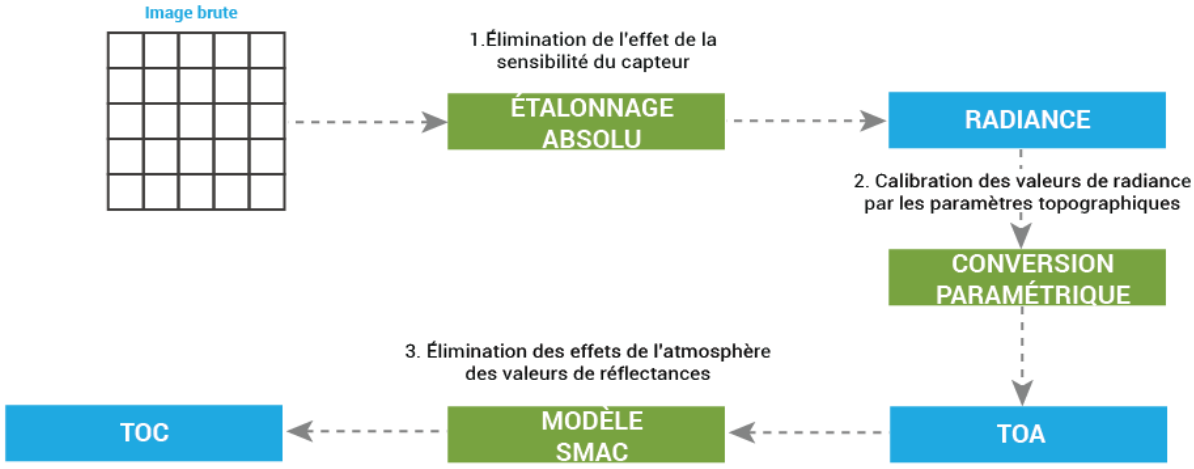


Figure 16 : Processus d'étalonnage des valeurs spectrales

D'après la littérature, les images satellites subissent également une altération due aux effets de l'atmosphère ce qui engendre une altération des valeurs radiométriques enregistrées par le capteur. Cette altération est la conséquence directe de deux phénomènes naturels dénommés l'absorption et la diffusion. En effet, l'absorption est le processus par lequel l'énergie électromagnétique change de forme. Tandis que, la diffusion représente le processus de dispersion de la lumière par la matière qui est dans les deux cas les gaz et les aérosols (Kergomard, Pratique des corrections atmosphériques en télédétection : utilisation du logiciel 5S-PC; 2000). Dans le but de remédier à ce problème, plusieurs méthodes et techniques ont été développées pour aboutir à cette fin, mais l'une des plus connues, c'est le modèle simplifié de correction atmosphérique SMAC (Rahman et Dedieu 1994) qui est un modèle de conversion de la réflectance au top d'atmosphère à une réflectance au sol. Dans le cadre de cette analyse, nous avons utilisé le modèle SMAC implémenté par Hagolle Olivier du CESBIO[42] sous Python sur lequel il existe un certain nombre de paramètres à définir et qui dépendent des mesures atmosphériques de la même date de la prise de vue. C'est donc pour

[42] Centre d'Etudes Spatiales de la BIOsphère

cette raison que nous avons utilisé les paramètres suggérés pour une correction approchée[43].

Calcul des indices spectraux

Suite à la phase de correction et de calibration optique des images et grâce à la variété spectrale du capteur des satellites Pléiades, nous avons calculé certains indices spectraux qui seront la base de discrimination entre les différents types d'occupation des sols. Nous avons choisi donc d'utiliser dans cette chaîne de traitement les indices les plus reconnus en matière d'efficacité. Ces derniers sont tous décrits ci-après:

NDVI[44] : l'indice le plus adapté à l'extraction de la végétation, il est calculé à partir de la bande rouge et infrarouge. Les valeurs de cet indice varient entre -1 et +1 (Mašková, Zemek et Květ 2008).
Le NDVI se définit par :

$$NDVI = \frac{PIR - R}{PIR + R}$$

Où
PIR : la réflectance en proche infrarouge.
R : la réflectance en rouge.

NIRR [45] : la réflectance très importante des zones végétalisées en proche infrarouge fait que la valeur de ce rapport entre la réflectance en infrarouge et les autres bandes réserve toujours les grandes valeurs pour la végétation (Jabari et Zhang 2013). La formule qui représente cette description est la suivante :

$$NIRR = \frac{PIR}{PIR + R + B + V}$$

Où
PIR : la réflectance en proche infrarouge.
R : la réflectance en rouge.
B : la réflectance en bleu.
V : la réflectance en vert.

NDWI[46] : un indice normalisé dérivée des bandes du proche infrarouge et du vert. Il permet la détection de l'eau en s'appuyant sur la bande infrarouge qui est caractérisée par sa capacité de mettre en contraste les surfaces végétalisées et les surfaces d'eau (McFeeters 1996).
Il s'exprime par la formule suivante :

$$NDWI = \frac{G - PIR}{G + PIR}$$

[43] www.cesbio.ups-tlse.fr
[44] Normalized Difference Vegetation Index
[45] Near Infrared Ratio
[46] Normalized Difference Water Index

Il existe également un autre indice portant le nom du NDWI et qui permet de décrire la teneur en eau de la végétation. Il est calculé à partir de la bande du moyen infrarouge et du proche infrarouge (Gao 1996).

$$NDWI = \frac{NIR - SWIR}{NIR + SWIR}$$

Où
PIR : la réflectance en proche infrarouge.
G: la réflectance en vert.
SWIR : la réflectance en moyen infrarouge.

L'indice utilisé dans cette analyse est celui de (McFeeters 1996) et non pas (Gao 1996) celui de car ce qui nous intéresse c'est la détection des surfaces en eau et non pas décrire la teneur en eau du couvert végétal.

BRIGHTNESS : la luminosité est définie comme étant la moyenne des différentes bandes spectrales de l'image (Jabari et Zhang 2013). Elle est exprimée par la formule ci-dessous :

$$BRIGHTNESS = \frac{R + G + B + PIR}{4}$$

Où B, V, R, IR sont les comptes numériques des différentes bandes spectrales.
NB : à l'issue de cette étape, tous les indices ont été concaténés avec les bandes de bases de l'image Pléiades pour ne former qu'un seul bloc d'informations.
INDICE DE BRILLANCE : un indice utilisé souvent en pédologie car il permet de discriminer entre plusieurs types de sols. Il est exprimé par la formule ci-après :

$$IB = \sqrt{R^2 + PIR^2}$$

Avec :
R et PIR les réflectances respectives de la bande rouge et proche infrarouge

Filtrage

Après avoir calculé les indices spectraux, nous avons entamé une opération permettant d'homogénéiser les valeurs radiométriques dans l'image en éliminant les bruits et réduisant les variations de la radiométrie qui fausserait éventuellement les prochains processus que nous allons opérer sur ces images, notamment la segmentation. Cette opération porte le nom du filtrage.
Plusieurs techniques existent pour le filtrage, mais chacune à un rôle spécifique qui diffère des autres types de filtres. Dans la chaîne de traitement et dans cette phase de prétraitement, les filtres sont utilisés dans le but d'éliminer les petits objets dans un premier temps et lisser l'image dans un second temps.

Le premier filtre utilisé pour atteindre ces objectifs est le filtre médian (Devarajan, Aatre et Sridhar 1900) qui représente un filtre non linéaire. Il se base sur une fenêtre de convolution sur laquelle il remplace la valeur de chaque pixel par la valeur médiane du voisinage de ce dernier. Dans le cas de ce travail, il a servi à éliminer les voitures et les objets de petite taille qui constituent un bruit de type sel et poivre (Salt and Pepper).

Comme seconde étape du filtrage, il a fallu faire appel à un autre filtre de famille des filtres non linéaires qui est le Meanshift. Il représente un lissage par moyennage qui préserve les contours comme le cas du filtre médian. En outre, il vise à faire converger chaque point vers son maximum local le plus proche, c'est-à-dire, qu'il cherche à stabiliser la valeur de chaque point par rapport à son voisinage[47].
L'algorithme de ce filtre applique le principe (Michel, Youssefi et Grizonnet 2015) décrit ci-après:

- Soit $p \in S_{pixels}$ où S_{pixels}L'ensemble des pixels de toute l'image.

Filtrage pour chaque pixel:

- Pour chaque p de S_{pixels}
  - Recherche d'un voisinage proche spectralement du pixel p.
  - Estimation de la moyenne spectrale du voisinage de p.
  - Réitération du processus jusqu'à ce que la valeur de p ne varie plus (Convergence).

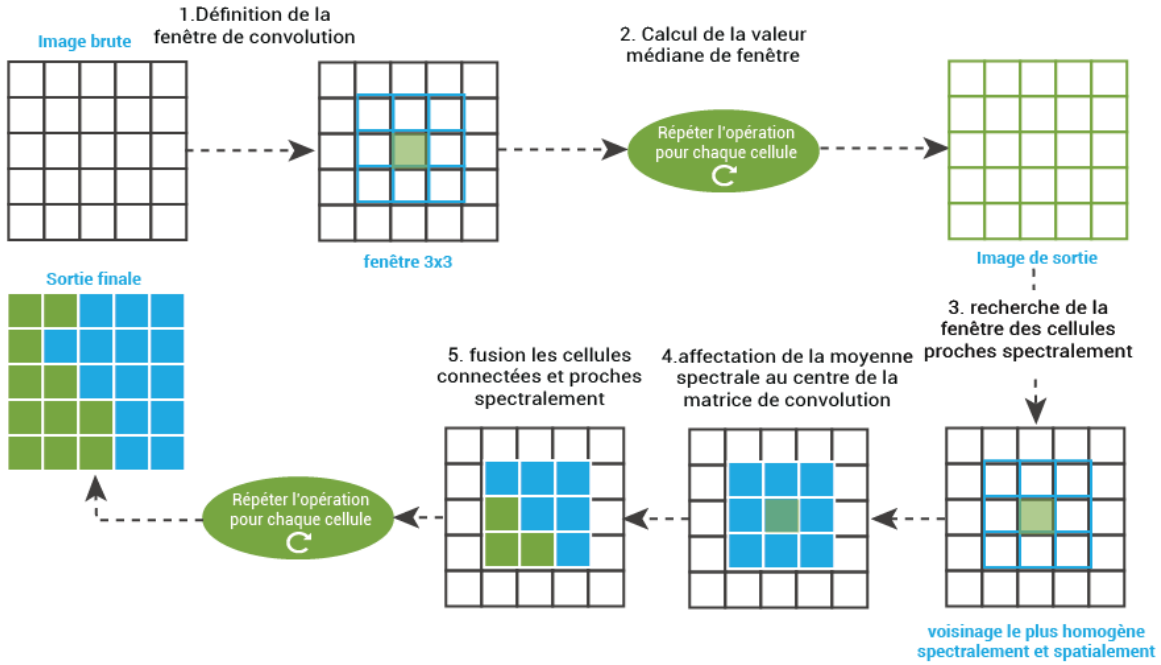


Figure 17 : Processus de filtrage

NB: la version utilisée du filtre Meanshift c'est celle qui est implémentée dans l'OrfeoToolBox par le CNES.

Classification orientée objet

L'approche orientée objet est une nouvelle méthode qui devient de plus en plus utilisée dans les travaux scientifiques de classification de l'occupation des sols, plusieurs travaux ont montré que cette approche offre des résultats plus intéressants que la classification pixel par pixel (Baatz et Schape 2000), (Kagamata, et al. 2005) surtout dans le milieu urbain comme (Bhaskaran, Paramananda et Ramnarayan 2010) le confirme dans sa publication. Cette approche particulière de classification se définit généralement comme étant une méthode qui se passe du pixel pour adopter une nouvelle échelle qui est le niveau objet. Ce

[47] xphilipp.developpez.com/articles/meanshift/

dernier est le regroupement d'un certain nombre de pixels ayant les mêmes caractéristiques dans un voisinage défini.
En effet, la classification orientée objet nécessite une donnée dotée d'une bonne résolution spatiale et prétraitée d'une manière à ce que les objets d'intérêt soient claires et non ambigus c'est pour cette raison que nous avons appliqué une phase de filtrage pour améliorer le rendu visuel de l'image d'entrée.
L'approche orientée objet est répartie sur deux étapes principales :
Segmentation : elle permet d'assurer le passage du niveau pixel au niveau objet par le biais d'un regroupement des pixels avec une similarité spectrale, textural et spatial au sein d'un même objet, ce qui permet la création d'un ensemble de zones homogènes. Il existe une multiplicité de méthodes permettant d'assurer cette opération parmi lesquelles, nous avons trouvé la méthode Meanshift implémentée sous OTB et qui représente la parfaite continuité pour les opérations de filtrage adoptées dans l'étape précédente.
Le paramétrage de la fonction LSMSMeanShift d'OrfeoToolBox a été réalisé d'une manière permettant d'éviter la sur-segmentation, c'est qui se traduit par le fait que la fonction regroupe les pixels de deux classes distinctes en une seule classe. Dans l'optique d'éviter ce problème, il a fallu réaliser plusieurs essais pour trouver le bon compromis entre le rayon spectral et le rayon spatial à fin d'éviter la sous-segmentation et la sur-segmentation. Finalement c'est l'essai 4 qui représente le bon ajustement des paramètres permettant d'éviter la sur-segmentation et la sous-segmentation.

Paramètres	*Essai 1*	*Essai 2*	Essai 3	Essai 4
Rayon spectral	10	50	10	30
Rayon spatial	5	5	30	5

Table 3 : les tests appliqués pour la définition des paramètres de segmentation

Classification : elle est la deuxième étape de l'approche orientée objet où chaque objet est issu de la segmentation est attribué à une classe selon ses attributs spectraux, texturaux, géométriques ou contextuels en suivant une méthode de classification bien précise. Il existe deux familles de méthodes pour classifier les images multispectrales que ça soit au niveau objet ou au niveau pixel, on parle des classifications supervisées et non supervisées.
Concernant la classification non supervisée, elle est un type de clustering permettant de définir sur un ensemble de données deux à deux comparable une partition qui respecte au mieux les ressemblances entre ses éléments[48] sans passer par des éléments d'entraînement des modèles de classification ce qui est le cas des méthodes supervisées où des éléments d'entraînement représentant la vérité-terrain sont introduits pour l'entraînement d'un modèle de classification.
Dans le cas de notre travail, les attributs géométriques n'ont pas été utilisés vu que les objets d'intérêt de cette étude n'ont pas les mêmes caractéristiques géométriques, prenons à titre d'exemple les routes et les parkings comme il est évident sur la figure ci-dessous, les routes sont généralement d'une forme allongée au contraire de certains parkings qui peuvent avoir une forme circulaire.

[48] maths.cnam.fr/IMG/pdf/GuenaelSlides.pdf

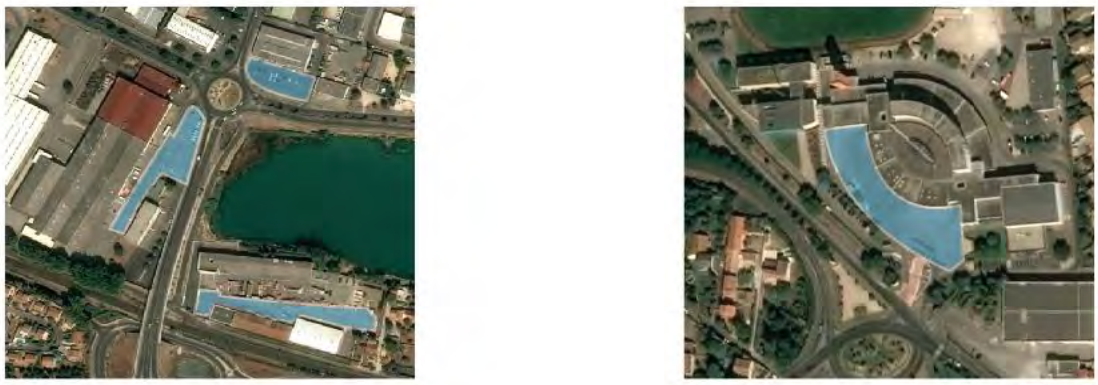


Figure 18 : Formes de quelques exemples de parkings

En effet, nous avons utilisé dans un premier temps une approche particulière pour la classification des objets obtenus suite à la segmentation, elle se base des règles de décision et elle porte le nom de RBC[49]. Cette méthode a été améliorée en introduisant un concept de la logique floue qui est le degré d'appartenance de chaque objet à une classe spécifique en partant des connaissances spectrales acquises de la littérature. Dans ce sens, 5 classes ont été définies (bâti, eau, végétation, surfaces artificialisées, ombre) et avec les fonctions d'appartenance définies pour chacun des indices spectraux.
Les fonctions d'appartenances utilisées pour les différents indices sont exprimées par les formules suivantes:

NDVI:

$$\mu_{NDVI} = \begin{cases} 1 & pour \quad 0 \leq x - T_2 \leq 1 - T_2 \\ 1 - \frac{T_2 - x}{T_2 - T1} & pour \quad 0 \leq T_2 - x \leq T_2 - T_1 \\ 0 & autrement \end{cases}$$

Avec T1 =0.19 et T2=0.3

NIRR:

$$\mu_{NIRR} = \begin{cases} 1 & pour \quad 0 \leq x - T_2 \leq 1 - T_2 \\ 1 - \frac{T_2 - x}{T_2 - T1} & pour \quad 0 \leq T_2 - x \leq T_2 - T_1 \\ 0 & autrement \end{cases}$$

Avec T1 =0.3 et T2=0.4

NDWI :

$$\mu_{NDWI} = \begin{cases} 1 & pour\ NDWI \geq T_1 \\ 0 & autrement \end{cases}$$

Avec T1 =0.15

[49] Rule Based Classification

BRIGHTNESS :

$$\mu_{brightness} = \begin{cases} 1 & pour \quad x < M \\ 1 - \frac{x-M}{3\sigma} & pour \quad 0 \leq x - M \leq 3\sigma \\ 0 & autrement \end{cases}$$

Avec M=310

Après la définition des fonctions à utiliser dans la création du degré d'appartenance, nous avons commencé par différencier l'ombre des autres objets en utilisant la luminosité. D'après la littérature (Shi et Li 2012), l'ombre est caractérisée par des valeurs de luminosité très faible, c'est pour cette raison qu'une centaine d'éléments de référence ont été pris pour déterminer la valeur de seuillage d'ombre à prendre en compte dans la fonction du degré d'appartenance à cette classe. Grâce à cette procédure, il a été remarqué que les objets représentant l'ombre leurs valeurs en compte numérique de luminosité sont inférieurs à 390, mais, il existe quelques exceptions que la fonction de degré d'appartenance utilisée est capable de gérer grâce à sa deuxième valeur de seuil ($T_2 = 3*\sigma$ avec σ l'écart type des valeurs numériques inférieur à 310), et sachant que les ombres sont des éléments opaques avec une faible variation radiométrique donc l'écart type ne sera pas très grand ce qui fait que la deuxième valeur de seuil sera très proche de la première.

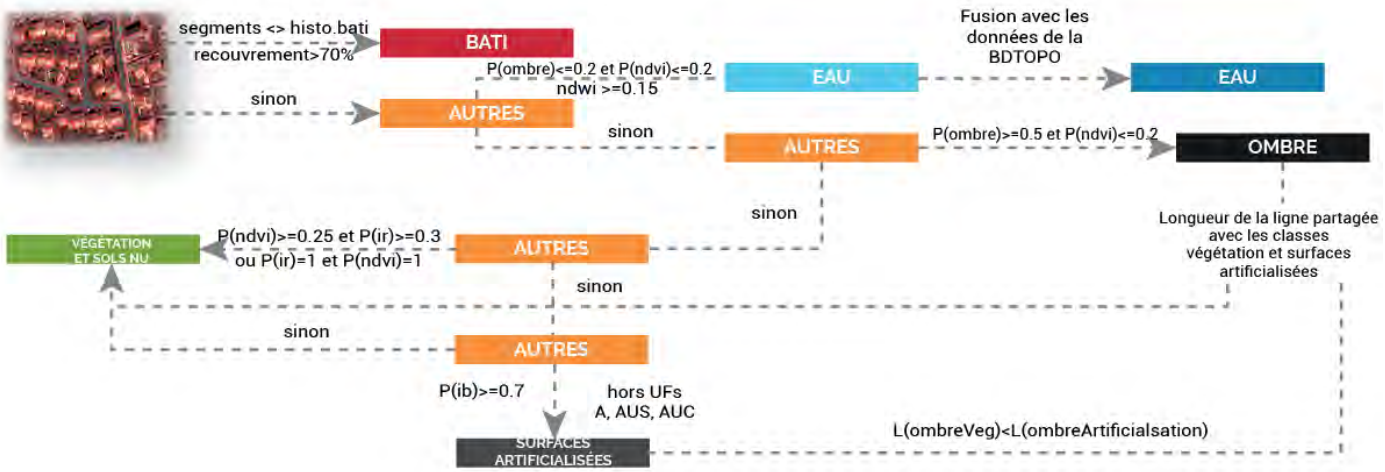


Figure 19 : Schéma de classification par un arbre de décision

Par la suite et concernant l'identification de la végétation deux indices ont été utilisés, le NDVI qui est l'indice de végétation normalisé et le NIR ratio qui permet la distinction de la végétation grâce à la forte réflectance de cette dernière dans la partie proche infrarouge du domaine spectral comme il a été annoncé dans le travail de (Jabari et Zhang 2013), les valeurs les plus grandes de cet indice sont considérées comme de la végétation. Les valeurs de végétation de ces deux indices sont supérieures à 0.19, mais avec quelques exceptions. C'est pour cela qu'on introduit encore une deuxième valeur de seuil et qui représente la valeur à partir de laquelle on est certain d'avoir la végétation (NDVI=0.3), pour estimer le degré d'appartenance à la classe « végétation et autres» avec le NDVI. La même fonction a été utilisée pour le reste des indices, mais avec des valeurs de seuils différentes.

Suite à l'estimation du degré d'appartenance, la classification a été débutée par une identification de la classe « bâti » en se basant sur l'existant à savoir une couche vectorielle stockée dans la base de données du projet URBANSIMUL qui modélise le bâti, elle a été croiser avec le résultat de l'étape de segmentation pour considérer les objets couverts par cette couche avec un pourcentage de recouvrement supérieur à 70% (recouvrement par rapport à la superficie totale de l'objet) comme étant du bâti.

Ensuite, une seconde règle a été appliquée sur les autres objets pour la détection des surfaces d'eau cette fois-ci en s'appuyant sur le degré d'appartenance à la classe « eau », calculé à l'aide du NDWI. D'après (McFeeters, Using the Normalized Difference Water Index (NDWIà within a Geographic Information System to Detect Swimming Pools for Mosquito Abatement: A Practical Approach 2013) l'eau en NDWI possède des valeurs positives en se basant sur cette idée, la valeur de seuil qui a été prise pour le NDWI c'est 0.15. En s'appuyant sur cette information et les degrés d'appartenance à la classe « ombre » et la classe « végétation et autres » par NDVI, la classe eau a été définie comme étant les objets qui possèdent une valeur de NDWI supérieure à 0.15 et une probabilité inférieure à 20% pour être de l'ombre ou de la végétation. Le résultat de cette condition a été par la suite fusionné avec l'existant à savoir la couche surface d'eau issue de la BD TOPO.
Après la définition de la classe « bâti » ainsi que la classe « eau » , une extraction d'ombre a été réalisée en se basant sur la probabilité que l'objet soit de l'ombre et qu'il ne soit pas de la végétation (P(ombre)>=0.5 et P(ndvi)<=0.2), cette opération a été vérifiée par la suite et il s'est avéré que les résultats se rapprochent de la réalité. L'extraction de l'ombre nous a laissé face à deux classes les surfaces artificialisées, la végétation et autres, l'extraction de ces deux derniers éléments nous l'avons réalisé grâce à la probabilité d'appartenance à la végétation. D'après (Lillesand et Kiefer 1994) la valeur du NDVI pour les sols nus peut se rapprocher des valeurs de la végétation, c'est donc pour cette raison que le degré d'appartenance prit en compte pour discrimination de la classe « végétation et autres » a été fixé à 0.25 pour le NDVI et 0.3 pour le NIR.
Après toutes ces étapes de classification, il a été remarqué qu'il reste les surfaces artificialisées à savoir les surfaces asphaltées, les espaces verts artificialisés, quelques surfaces brillantes du voisinage du bâti et des éléments du sol nu. Le problème rencontré dans cette partie résidait dans le fait que nous étions dans l'incapacité de distinguer entre les sols nus restants et les autres classes en absence d'une bande SWIR où il est possible de mettre en évidence les sols nus grâce à l'indice BSI (bare soil index). Face à cette situation, l'aide de l'information existante nous a encore une fois été d'une grande utilité, car grâce au zonage du PLU (Plan Local d'Urbanisme) nous avons pu définir les unités foncières qui sont dans des zones à bâtir ou dans une zone agricole pour éliminer les sols nus et les terres labourées qui apparaissent dans ces types de zonage. En outre, l'élimination des surfaces brillantes a été réalisée grâce à l'indice de brillance par le même principe que la luminosité avec seuil qui égale à 0.25.
Quant à la classe « ombre », elle a subi un traitement spécifique pour pouvoir la classer entre la classe « végétation et autres » et la classe surfaces artificialisées. Dans le travail de (Singh et al, 2012), l'ombre est définie comme une altération de l'information générée par les objets élevés comme les bâtiments et les arbres. En partant de ce constat nous avons considéré que tout objet de type ombre est considéré comme étant un élément de la classe « végétation

et autres » si et seulement si la géométrie de l'objet touche une géométrie de la classe « végétation et autres » sinon s'il croise aussi la classe « surfaces artificialisées », la longueur de la ligne partagée entre la classe « ombre » et les deux autres classes est calculée pour attribuer la classe de l'objet qui partage la plus longue ligne avec classe « ombre » à l'objet ombre en question. Et finalement si la géométrie de l'ombre ne touche que les objets de la classe « végétation et autres » cette dernière est considérée comme classe de cette ombre.

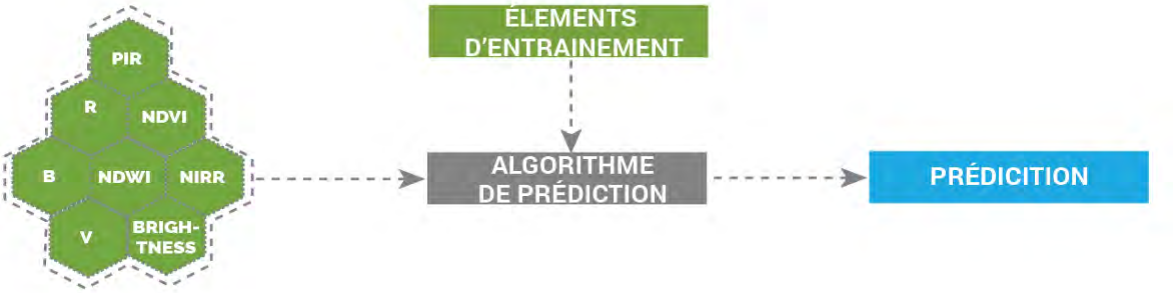


Figure 20 : Schéma explicatif de la procédure de classification par les modèles automatiques

Dans un second temps, nous avons opté pour l'utilisation des méthodes de classification supervisée automatiques considérées comme étant plus robustes que la classification suivant des règles de décisions (Ma, et al. 2017).
Dans l'optique d'évaluer l'apport de la classification via les méthodes automatiques par rapport à la classification précédemment décrite deux algorithmes ont été choisis.

Le premier est le Random Forest ou Forêt aléatoire, il représente une combinaison de plusieurs modèles de type arbre de décision qui cherche à prédire la classe de chaque objet en utilisant un grand nombre d'arbres de décision construits, chacun entraîné à partir d'un sous-ensemble issu de l'ensemble d'apprentissage original en introduisant le concept de randomisation (Breiman 2001). L'algorithme Random Forest est caractérisé par sa résistance au sur-apprentissage vu que les arbres de décision convergent dans le cas où il existe un nombre suffisant d'arbres (Genuer, Poggi et Tuleau 2008). Le second c'est le SVM ou séparateurs à vaste marge un algorithme de discrimination basée sur la détermination de l'hyperplan qui sépare les éléments des différentes classes tout en maximisant la distance entre ces dernières. Cette maximisation de marge procure plus de sécurité et minimise l'erreur (Desir 2013).

Ces algorithmes de classification automatique se basent principalement sur des entrées de référence permettant l'apprentissage du modèle à fin de prédire la classe de chaque objet. Dans le cas de notre travail, la définition des éléments de référence a été réalisée d'une manière automatique et aléatoire en se basant sur les règles de décision de la première méthode décrite précédemment avec l'élimination de l'ombre, car ce dernier faussé les résultats des prédictions lors des premiers tests des méthodes de classification automatiques. En effet, le choix des seuils pour cette méthode d'entraînement automatisée s'est basé sur la littérature (Jabari et Zhang 2013) avec quelques éléments d'amélioration supplémentaires qui sont issus de l'analyse de la donnée utilisée. Pour chaque classe on récupère le maximum d'objets dans la limite de 2000 objets par classe qui respectent les

règles définies et d'une manière aléatoire, cela va permettre au modèle de minimiser le biais de la prédiction.

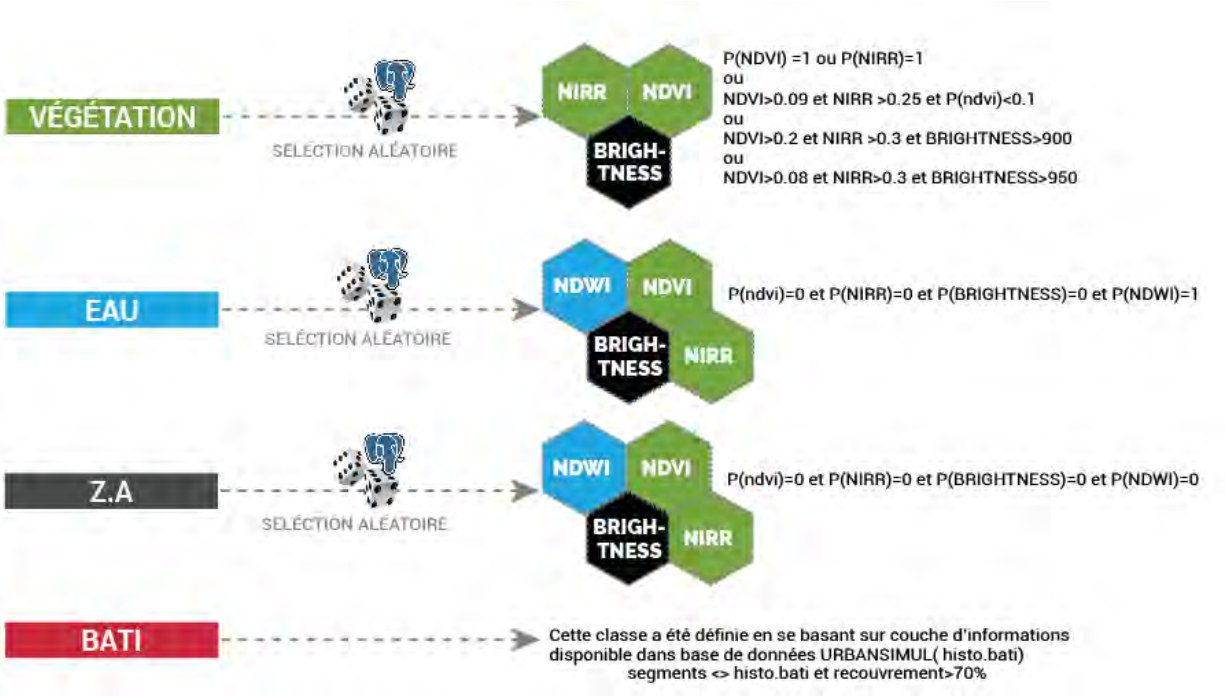


Figure 21 : les règles de décision utilisées pour l'apprentissage des modèles de classification

Validation : après la classification des objets issue de la phase de segmentation, une validation est nécessaire pour estimer la fiabilité des résultats, l'une des méthodes les plus utilisées pour aboutir à cette fin est l'indice de Kappa qui mesure l'accord entre deux variables qualitatives ayant les mêmes modalités (Santos 2017). Ensuite, une seconde étape de validation a été utilisée qui est la validation par photographie aérienne de Bing Maps [50] dans le but de vérifier de plus près les résultats obtenus.

Post-traitements : dans le but d'améliorer le rendu de la géométrie, un filtre de lissage a été appliqué sur les surfaces artificialisées détectées. Ce filtre porte le nom de la fermeture, il est le fruit d'une combinaison de deux autres filtres (dilatation et érosion).
NB : Le filtre fermeture a été appliqué en utilisant une fenêtre de 7x7 avec l'implémentation disponible avec la bibliothèque Scipy[51].

c. Résultats et discussions

L'application des méthodes développées pour la classification des images Pléiades d'une manière générale et l'extraction des surfaces artificialisées particulièrement à travers la détection orientée objet a donnée des résultats de précision variable. C'est dans cette partie

[50] www.bing.com
[51] docs.scipy.org

que des exemples de sorties vont être exposés et discutés en fonction de chaque étape de la chaîne de traitement élaborée.

i. Prétraitements

Les images acquises dans le cadre du programme GEOSUD ont subi quelques opérations permettant dans un premier temps d'ajuster les valeurs radiométriques grâce à la correction radiométrique et atmosphérique telle qu'elles sont décrites dans la partie méthodologie. Ces opérations sont primordiales dans le calcul des indices spectraux, par exemple et comme il est évident dans la figure 25 la valeur de l'indice de végétation normalisée pour une même surface a connu une augmentation.

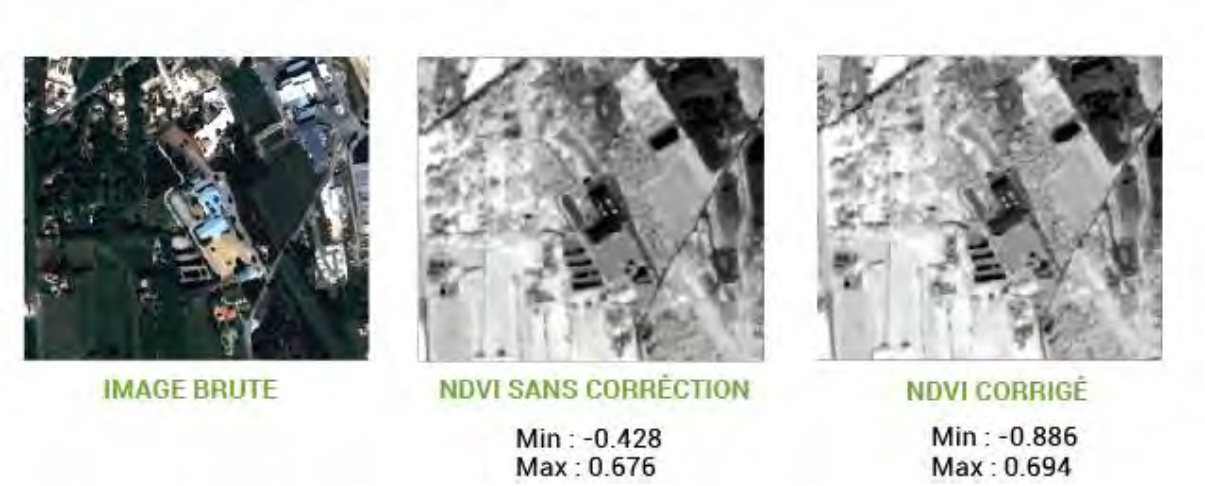


Figure 22 : Comparaison visuelle entre les valeurs de l'NDVI avant et après la correction

Après la correction radiométrique et atmosphérique appliquée sur les images brutes, un lissage par filtre médian a été d'une grande nécessité pour éliminer les bruits de type poivre et sel. Ce dernier a été suivi par un lissage par filtre Meanshift à fin d'homogénéiser les valeurs radiométriques.
À travers la Figure 26, nous pouvons voir ce qui a été énoncé précédemment sur le rendu des deux types de filtres. On constate que le filtre médian, avec les paramètres pris en compte, efface les objets avec une faible dimension spatiale tels que les voitures et les traits de grandes routes. Par rapport au filtrage Meanshift, il est évident sur la Figure 26 qu'il stabilise la valeur radiométrique dans un voisinage donné caractérisé par une ressemblance spectrale.

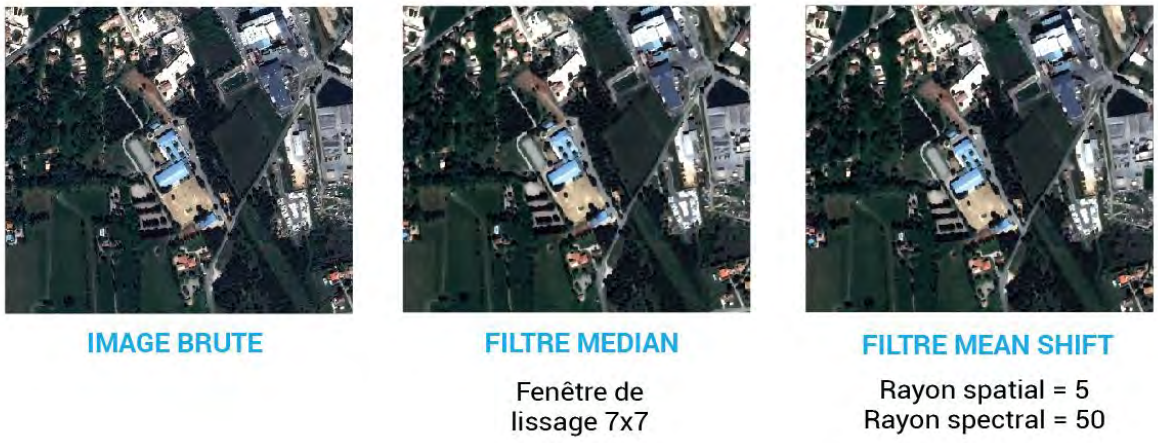


Figure 23 : Différence entre l'image brute et les images filtrées

En parallèle à la phase du filtrage, le calcul des différents indices spectraux a été réalisé ainsi qu'une concaténation des résultats de ce dernier calcul avec les images brutes dans le but d'obtenir qu'une seule entrée pour l'étape suivante à savoir la classification orientée objet.

ii. Classification

Avant d'entamer la classification, la segmentation a été réalisée pour transformer les pixels en objets par la méthode Meanshift comme il est apparent sur la figure ci-dessus. On constate que la segmentation offre plusieurs objets caractérisant différents types d'occupation du sol, c'est donc pour cette raison que nous avons appliqué la classification de ces objets de manière rigide dans un premier temps avec les règles de décision, avant d'adopter au final les méthodes automatiques de classifications.

Indicateurs	*RBC**	*SVM*	RF
Indice de Kappa	75.41	83.3	99.19
Accord global	75.45	83.1	99.71

Table 4 : la précision des modèles de classifications utilisés

Ces deux méthodes décrites précédemment dans la partie méthodologie ont été évaluées grâce à l'indice de Kappa, qui nous a montré que les algorithmes utilisés donnent des résultats très intéressants, mais avec des niveaux de précision variables. Pour le cas de la première méthode qui est la RBC[52], son évaluation a donné un indice de Kappa supérieur à 75%, cela veut dire que les 75% des résultats sont acceptés, mais ce pourcentage reste inférieur à celui des méthodes de prédiction.
En effet, même au sein de cette famille de modèles de prédiction, l'écart est très important, car, l'évaluation du modèle SVM nous a montré que ce dernier reste moins puissant que le modèle Random Forest dans le cas de cette étude caractérisée par le nombre important de variables prises en compte lors de cette phase de classification. L'indice de kappa pour

[52] RBC : Rule Based Classification

l'exemple du Random Forest il dépasse 99% cette valeur indique que plus de 99% des éléments classifiés correspondent à la vérité terrain sur un grand nombre d'entités d'entraînement même si la méthode s'appuie sur le principe de randomisation qui reste un point fort offrant une grande résistance à ce modèle au sur-apprentissage dans ce genre de situation (Girard et Girard 1999) .
Ci-dessous la matrice de confusion du modèle Random Forest:

		CLASSIFICATION				
		VÉGÉTATION ET AUTRES	EAU	SURFACES ARTIFICIALISÉES	BATI	PRÉCISION RÉALISATEUR
TERRRAIN	VÉGÉTATION ET AUTRES	576405	5	306	3	99.94
	EAU	363	39553	1133	61	96.21
	SURFACES ARTIFICIALISÉES	1	0	63190	207	99.67
	BATI	0	0	0	44666	100
	PRÉCISION UTILISATEUR	99.93	99.98	97.77	99.39	

Indice de Kappa : 99.19% Accord global : 99.71%

Table 5 : Matrice de confusion pour l'ensemble des dalles choisies pour la chaîne de traitement

La validation des résultats de la classification réalisée s'est appuyée également sur les photos aériennes de Bing Maps pour vérifier la pertinence des entités géographiques obtenues par les différentes méthodes. Ces résultats se sont avérés satisfaisants et répondent en grande partie aux besoins du projet URBANSIMUL avec quelques imperfections.

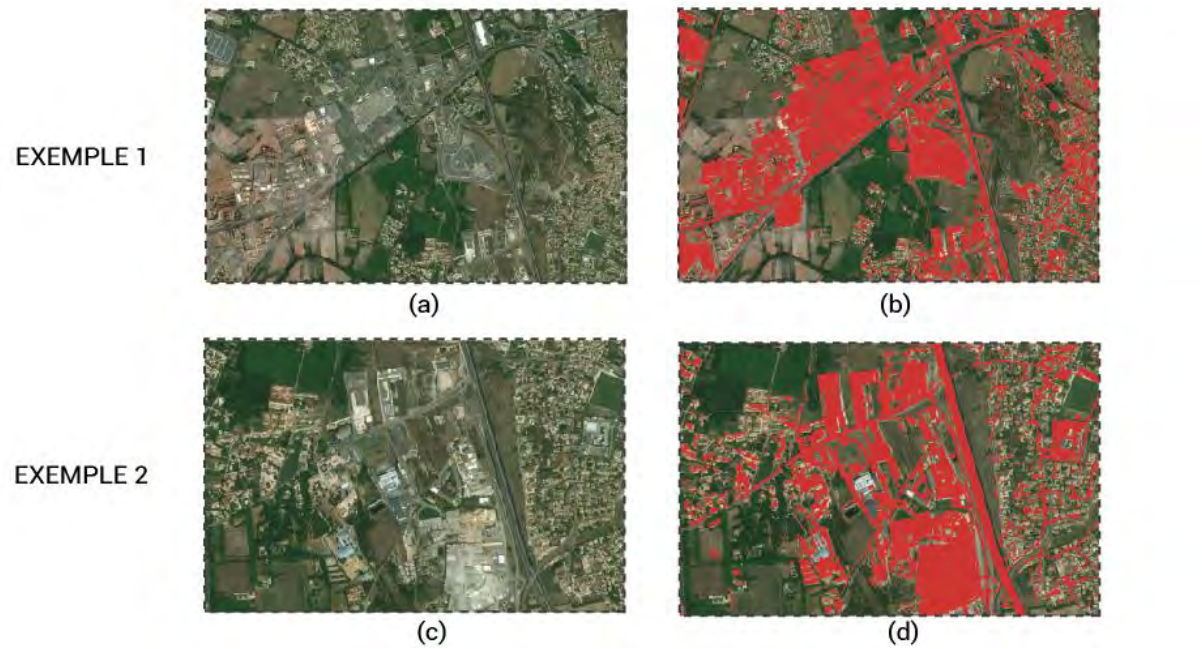


Figure 24 : Exemples des prises aériennes de validation

Cette validation nous a montré que les résultats sont généralement bons en termes de prédiction du type d'occupation du sol et de précision géométrique des entités

géographiques représentées, notamment pour la classe des surfaces artificialisées. Il a été remarqué que plusieurs zones artificialisées ne sont pas couvertes par la couche issue de l'application de la classification par les règles de décisions, chose qui peut être expliquée par la rigidité des règles posées. En revanche, les méthodes de prédiction surtout la méthode Random Forest a donné des bons résultats qui se rapproche le plus de la réalité malgré, le fait qu'elle confond entre les zones artificialisées et quelques sols labourés. Cependant, les résultats de Random Forest restent quand même mieux que ceux de SVM où il a été remarqué que ce dernier subit le problème de sur-apprentissage provoqué par l'incapacité de ce modèle à traiter un nombre important de classes, car à la base il est meilleur en prédiction binaire (Hsu et Lin 2002).

Enfin, ce sont les résultats obtenus par la méthode de classification Random Forest qui ont été gardés pour pouvoir généraliser la méthodologie sur l'ensemble de la région PACA et intégrer cette couche d'informations dans les bases de données du Projet URBANSIMUL.
Ci-après un petit aperçu du résultat final de la chaîne de traitement autonome développée dans le cadre de ce stage :

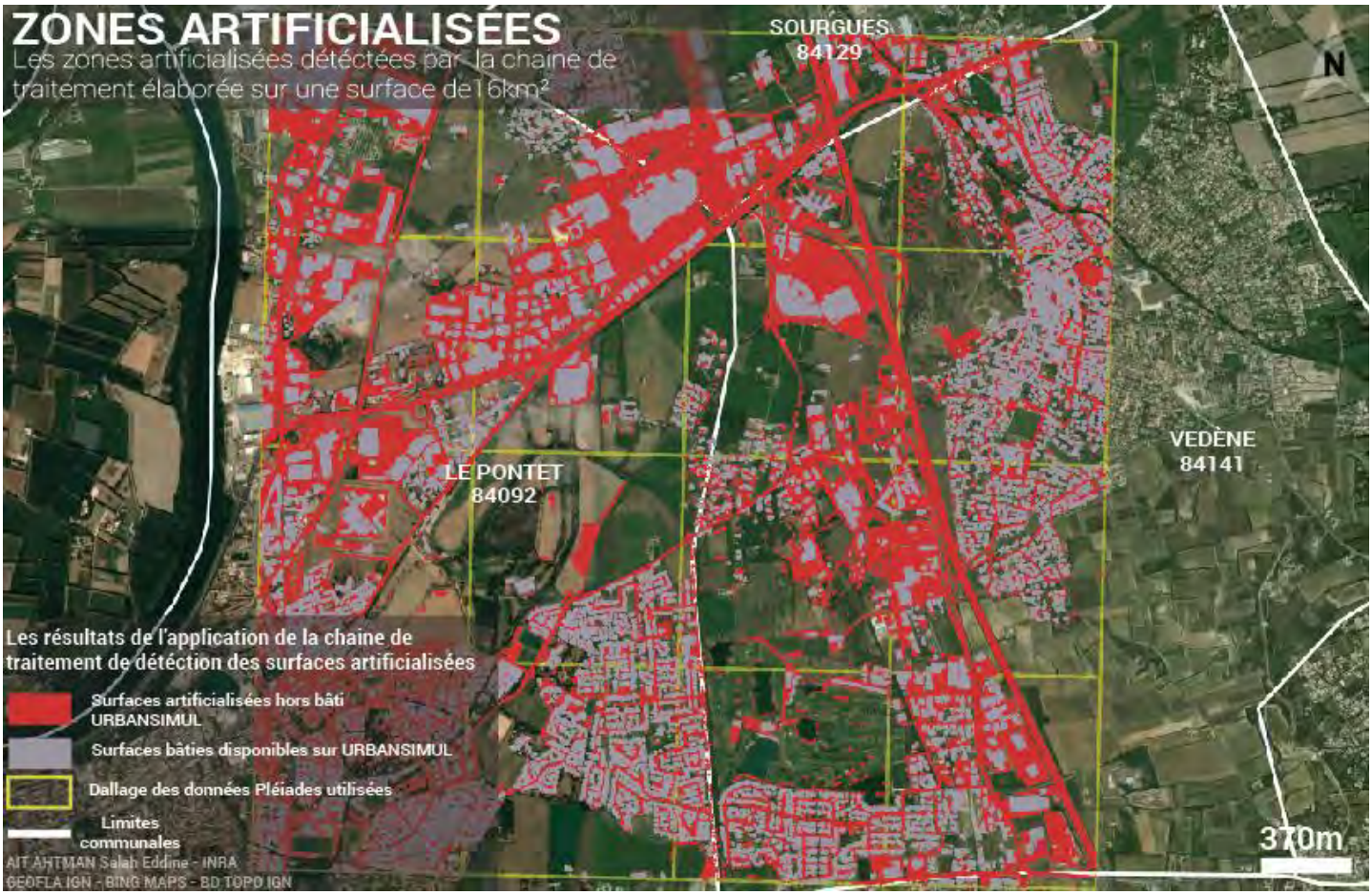


Figure 25 : Cartographie des surfaces artificialisées détectées sur la zone des tests

BILAN ET PERSPECTIVES

1. Conclusion

Dans la région PACA, la gestion du territoire est d'une grande complexité dans la présence des phénomènes comme l'étalement urbain et la périurbanisation. Au cours des dernières années, ces phénomènes ont évolué d'une façon remarquable caractérisée par le déséquilibre généré entre l'offre et la demande dans le marché foncier de la région. Dans l'optique de remédier à ces problèmes, la région PACA mise sur les établissements de recherche et d'étude de développement urbain comme l'INRA et le CEREMA pour trouver des solutions d'analyse et de simulation de l'offre foncière.
L'outil URBANSIMUL permet en réalité de répondre en grande partie aux besoins de la région en ce qui concerne la gestion du foncier, et ce travail rentre dans le cadre de l'évolution que connaissent cet outil et les données sur lesquelles il se base pour offrir les analyses et le rendu de disponibilité foncière.

Dans ce stage, tout d'abord, nous avons commencé par vérifier la pertinence des données libres d'OpenStreetMap. Cette source est une vraie mine d'informations, car elle s'est avérée d'une qualité plus ou moins bonne par rapport à la BD TOPO. L'évaluation des données OpenStreetMap a été réalisée grâce à deux types d'indicateurs, le premier, c'est l'indicateur d'exhaustivité et le second, c'est l'indicateur de précision géométrique. Ces derniers nous ont prouvés que cette source peut être complémentaire à la BD TOPO dans le cadre du projet URBANSIMUL, et particulièrement l'extraction des surfaces artificialisées. Dans un second temps, nous avons exploré la piste du traitement des images satellites à très haute résolution pour caler une chaîne d'identification des surfaces artificialisées. Cet objectif a été atteint en partie vu que la méthode élaborée a donnée des résultats caractérisés par une bonne précision suite à la validation effectuée à l'issue du traitement d'identification, mais les résultats sont également caractérisés par quelques imperfections à savoir la confusion entre quelques types de sols nus et les surfaces artificialisées.

Afin de permettre à l'équipe URBANSIMUL de reproduire les calculs des indicateurs de qualité pour les données OSM et l'extraction des zones artificialisées pour une zone donnée, un ensemble de scripts python ont été développés pour faciliter la reproduction des résultats de ce stage. Ces scripts seront améliorés par la suite pour optimiser les temps de calcul et le rendu des méthodes, particulièrement la confusion entre les surfaces artificialisées et les sols nus qui se rapprochent spectralement des surfaces artificialisées.

2. Retour d'expérience

Ce stage peut être qualifié comme étant une expérience très enrichissante, car il m'a permis à la fois de mettre en application mes acquis dans l'analyse de l'information géographique et la télédétection, ainsi que l'acquisition de nouvelles compétences. En effet, la première partie de ce stage m'a offert la possibilité de se familiariser avec une multitude de données et techniques que je n'ai jamais utilisées auparavant. Pour la seconde partie, elle m'a été d'une grande importance dans la découverte d'un nouvel outil de traitement d'images satellites, à savoir l'OrfeoToolBox et sa large palette de fonctionnalités.
Au-delà de l'aspect technique, ce stage a été pour moi une occasion de découvrir le milieu de la recherche d'une autre perspective, car le projet URBANSIMUL est bien au-delà, de la définition habituelle d'un simple projet de recherche. Cette expérience a été un véritable challenge pour moi vu que ce genre de méthodologie n'a pas été beaucoup développée et particulièrement dans le cas des images Pléiades. Grâce à l'encadrement de mon tuteur de stage GENIAUX Ghislain, mon tuteur enseignant FAUVEL Mathieu et sans oublier notre principal collaborateur LEROUX Bertrand du CEREMA que je ne saurais trop remercier, j'ai réussi à répondre aux attentes de ce stage en grande partie.
Enfin, j'ai pu présenter lors de la journée des stagiaires de l'unité qui a eu lieu le 19 juin 2017, l'avancement de mon stage et les premiers résultats que j'avais obtenus, en présence de la majorité de l'effectif de l'unité Ecodéveloppement ainsi que certains collaborateurs extérieurs. Cette présentation a eu une grande appréciation de la part des agents de l'unité répartis entre géomaticiens, sociologues, agronomes, économistes et informaticiens. Cela n'a pas été une chose facile vu que la plupart ne connaissent pas les fondamentaux de la gestion d'information géographique et la télédétection également, mais ce fut une agréable expérience.

BIBLIOGRAPHIE

Auber, Martin, Pierrick Billon, et Ophélie Petit. «les données routières d'OpenStreetMap dans la Sarthe : comparaison avec le RGE et contribution au projet.» Rapport universitaire, 2012.

Baatz, Martin, et Arno Schape. «Multiresolution Segmentation: an optimization approach for high quality multi-scale image segmentation.» 2000: 12-23.

Baley, Matthieu, et Guillaume Touya. «Intégration et correction automatique de données OpenStreetMap.» *Géomatique Expert*, 2014: 32-43.

Bhaskaran, Sunsil, Shanka Paramananda, et Maria Ramnarayan. «Per-pixel and object-oriented classification methods for mapping urban features using Ikonos satellite data.» *Applied Geography*, 2010: 650-665.

Breiman, Leo. «Random forests.» *Machine learning*, 2001: 5-32.

Desir, Chesner. «Classification automatique d'images, application à l'imagerie du poumon profond.» Thèse de doctorat, 2013.

Devarajan, Ganesh, Vasudev Kalkunte Aatre, et Sridhar. «Analysis of median filter.» *IEEE*, 1900: 274-276.

Genuer, Robin, Jean-Michel Poggi, et Christine Tuleau. «Random Forests: some methodological insights.» *arXiv*, 2008.

Girard, Michel-Claude, et Colette-Marie Girard. *Traitement des données de télédétection-2e éd.: Environnement et ressources naturelles.* Dunod, 1999.

Girres, Jean-François, et Guillaume Touya. «Quality assessment of the French OpenStreetMap dataset.» *Transactions in GIS*, 2010: 435-459.

Haklay, Mordechai. «How good is OpeenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England.» *Environnment & Planning B*, 2010: 682-703.

Hamilton, Randy, Kevin Megwon, Tom Mellin, et Ian Fox. «Guide to automated stand delineation using image segmentation.» *US Department of Agriculture, Forest Service, Remote Sensing Applications Center, Salt Lake City, Utah*, 2007.

Hsu, Chih-Wei, et Chih-Jen Lin. «A comparioson of methods multiclass support vector machines.» *IEEE transactions on Neural Networks*, 2002: 415-425.

Jabari, Shabnam, et Yum Zhang. «Very High Resolution Satellite Image Classification Using Fuzzy Rule-Based Systems.» *Algorithms*, 2013.

Kagamata, Noritoshi, Yukio Akamatsu, Masaru Mori, Yun Qing Li, Yoshinobu Hoshino, et Keitarou: Hara. «Comparison of pixel-based and object-based classifications of high resolution satellite data in urban fringe areas.» Hanoi: Asian Conference on Remote Sensing, 2005.

Kergomard, Claude. «La télédétection aéro-spatiale : une introduction.» 1990.

Kergomard, Claude. «Pratique des corrections atmosphériques en télédétection : utilisation du logiciel 5S-PC;.» *Cybergeo : European Jornal of Geography;*, 2000.

Lillesand, Thomas, et Ralph Kiefer. *Remote sensing and image interpretation.* Toronto: John Wiley and Sons, 1994.

Ma, Lei, Manchun Li, Xiaoxue Ma, Liang Cheng, Peijun Du, et Yongxue Liu. «A review of supervised object-based land-cover image classification.» *ISPRS Journal of Photogrammetry and Remote Sensing*, 2017: 277-293.

Maboudi, Mehdi, Jalal Amini, et Michael Hahn. «Objetcs grouping for segmentation of roads network in high resolution images of urban areas.» *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2016: 897-902.

Mašková, Zuzana, František Zemek, et Jan Květ. «Normalized difference vegetation index (NDVI) in the management of mountain meadows.» *Boreal environment research*, 2008.

McFeeters, Stuart. «The use of he Normalized Difference Water Index (NDWI) in the delieation of open water features.» *International journal of remote sensing*, 1996: 1425-1432.

McFeeters, Stuart. «Using the Normalized Difference Water Index (NDWIà within a Geographic Information System to Detect Swimming Pools for Mosquito Abatement: A Practical Approach.» *Remote Sensing*, 2013: 3544-3561.

Michel, Julien, David Youssefi, et Manuel Grizonnet. «Stable Mean-Shift Algorithm and its Application to the Segmentation of Arbitarily Large Remote Sensing Images.» *IEEE Transactions on Geoscience and Remote Sensing*, 2015: 952-964.

Pacifici, Fabio, Marco Chini, et William J. Emery. «A neural network approach using multi-scale textural metrics from very high-resolution panchromatic imagery for urban landuse classification.» *Remote Sensing of Environment*, 2009: 1276-1292.

Petit, Ophélie, Pierrick Billon, et Jean-Michel Follin. «Évaluation de la qualité des données OpenStreetMap sur la Sarthe et réflexion sur le processus de contribution.» *XYZ*, 2012: 24-34.

Qian, Yuguo, Weiqi Zhou, Jingli Yan, Weifeng Li, Han, et Lijian Han. «Comparing Machine Learning Classifiers for Obect-Based Land Cover Classification Using Very High Resolution Imagery.» *Remote Sensing*, 2015: 153-168.

Rahman, Hafizur, et Gérard Dedieu. «SMAC: a simplified method for the atmospheric correction of satellite measurments in the solar spectrum.» *Remote Senseing*, 1994: 123-143.

Santos, Frédéric. « Le kappa de Cohen: un outil de mesure de l'accord inter-juges sur des caractères qualitatifs.» 2017.

Shi, Wenxuan, et Jie Li. «Shadow detection in color aerial images based on HSI space and color attenuation relationship.» *EURASIP Journal on Advances in Signal Processing*, 2012: 141.

Viry, Mathieu, Thimothée Giraud, Marianne Guérois, Ronan Ysebaert, Nicolas Lambert, et Amel Feredj. *Généralités liées à OpenStreetMap et la complétude des données.* Rapport technique, CGET, 2016.

Wang, Ming, Qingquan Li, Qingwu Hu, et Meng Zhou. «Quality analysis of OpenStreetMap map data.» *International archives of the photogrammery, remote sensing and spatial information sciences*, 2013.

Zielstra, Dennis, et Alexander Zipf. «A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany.» 2010.

TABLE D'ILLUSTRATIONS

TABLE DES TABLEAUX