Télédétection et épidémiologie en zone urbaine : de l'extraction de bâtiments à partir d'images satellite à très haute résolution à l'estimation de taux d'incidence

Erika Upegui Cardona

UNIVERSITE DE FRANCHE-COMTE

ECOLE DOCTORALE « LANGAGES, ESPACE, TEMPS, SOCIETES »

Thèse en vue de l'obtention du titre de docteur en

GEOGRAPHIE ET AMENAGEMENT DES TERRITOIRES

TELEDETECTION ET EPIDEMIOLOGIE EN ZONE URBAINE

De l'extraction de bâtiments à partir d'images satellite à très haute résolution à l'estimation de taux d'incidence

Présentée et soutenue publiquement par

Erika UPEGUI CARDONA

Le 8 octobre 2012, à Besançon

Sous la direction de M. le Professeur Jean-François VIEL
et la codirection de M. le Directeur de recherche Daniel JOLY

Membres du Jury :

Jacques GARDON, Directeur de recherche à l'IRD, Montpellier, Rapporteur
Daniel JOLY, Directeur de recherche CNRS à l'université de Franche-Comté
Catherine MERING, Professeur à l'université de Paris Diderot (Paris 7)
Jean-François VIEL, Professeur à l'université de Rennes 1, Praticien Hospitalier
Christiane WEBER, Directeur de recherche CNRS à l'université de Strasbourg, Rapporteur

A ma mère, Gloria, pour m'avoir appris que « celui qui ne lâche pas, réussit »
A Javier pour m'avoir aimé malgré la distance

A mi madre, Gloria, por haberme enseñado que "él que persevera, alcanza"
A Javier por haberme querido a pesar de la distancia

REMERCIEMENTS

Ce qui a commencé, il y a quatre ans, comme une aventure motivée par ma passion par la télédétection a débouché sur une thèse. Ainsi, un jour d'été j'ai quitté mon pays avec le rêve de connaitre une autre culture, et d'approfondir mes connaissances dans le monde de l'imagerie satellitaire. J'ai traversé l'Atlantique et j'ai débarqué à Paris pour une année, le temps de faire le master en télédétection et géomatique appliquées à l'environnement. J'ai tellement aimé mon séjour que j'ai voulu rester pour continuer la découverte et l'apprentissage de nouvelles choses. Aussi, je tiens à remercier mon directeur de thèse, le professeur Jean-François VIEL pour la confiance qu'il m'a témoignée en me sélectionnant pour continuer ses recherches transdisciplinaires en télédétection et épidémiologie ; sa rigueur et sa disponibilité ont été fondamentales pour faire avancer cette recherche. Je tiens à exprimer ma profonde gratitude et reconnaissance à Catherine MERING qui a accepté ma candidature pour le master à Paris, me permettant ainsi de commencer cette aventure. Je la remercie aussi pour avoir soutenu ma candidature à cette thèse, ainsi que pour son soutien pendant la réalisation de cette dernière. Merci de m'avoir fait profiter de votre grande expérience.

Je tiens à remercier Daniel JOLY pour la confiance qu'il m'a accordée, en acceptant d'être co-directeur de cette thèse. Je remercie également Christiane WEBER et Jacques GARDON pour avoir consenti à être les rapporteurs de ce travail.

Pendant mes travaux, j'ai été accueillie à l'hôpital Saint Jacques, dans l'ancien département d'information médicale. Je tiens à remercier Fréderic MAUNY pour sa disponibilité, sa gentillesse, son expérience et son soutien tout au long de mon séjour. J'exprime ma gratitude à tous les membres du service, qui m'ont apporté leur sympathie, Anne, Anne-Lise, Emmanuelle, Eric, Fabienne, Laurence, Marc, Nicole, Sandrine, Sophie, Marie-Hélène, …, et tous les autres. J'ai eu également le plaisir de partager avec les internes en Santé Publique qui m'ont accueilli et m'ont fait participer à d'autres activités en dehors de la vie de l'hôpital : un grand merci, Annick, Ariane, Marie-Caroline, Maxime, Michel, Thomas et Rauchan.

Au cours de ces trois années, mes travaux de thèse m'ont amenée à travailler à l'Université Paris 7, au Pôle Image où j'étais toujours la bienvenue. Merci à toutes et à tous pour les « coups de main », pour votre soutien. Merci aux doctorants, post-docs et

aux stagiaires du laboratoire qui à un moment ou à un autre ont pu m'apporter leur aide. Merci pour les moments agréables, Ababakar, Benoît, Cecilia, Chantal, Claudia, François, Johanna, Joevin, José, Milena, Olivier, Oumar, …, et tous les autres.

J'exprime ma reconnaissance et ma gratitude envers toutes les personnes qui dans un moment ou un autre, m'ont fait partager leurs compétences pendant ces années de thèse. Merci à Antoine pour IDL, à Olivier pour R, à Quentin pour la biblio, à Marc pour la statistique, à Rami pour la programmation.

Je tiens à exprimer ma profonde gratitude et reconnaissance à Paul-Henri SCHMITT, professeur du CLA, pour la patience avec laquelle il a écouté mes nombreux doutes au sujet du français, et pour m'avoir fait partager son point de vue avec générosité. Je le remercie aussi pour le temps dédié à la lecture et à la correction de ce manuscrit. Toutes mes pensées amicales pour mes collègues du cours de FLE, Caroline, Erik, Jose Erlindo, Macei, Mei, Nicole, Rami, Sandra, et tous les autres.

Un grand merci aux personnels des bibliothèques : à l'Académie de Sciences, aux bibliothèques municipales de Besançon, ainsi qu'à l'Université Franche-Comté ; et également aux personnels de l'UFR SHLS, de l'ED LETS, de Chrono-environnement et de la MSHE.

Un grand merci à tous ceux qui m'ont fait confiance et m'ont aidé dans le recueil des données de la Colombie, malheureusement elles n'ont pas été suffisantes pour les inclure dans ce manuscrit. Merci aux personnels de l'IGAC, de l'EAAB., de la Secretaría Distrital de Salud de Bogotá et de DATUM Ingeniería. Mes remerciements s'adressent également à Angelica ACERO, Jorge CASTILLO, Andres DIAZ, Felipe FONSECA, Mauricio FUENTES, Meyra del mar FUENTES, John GUZMAN et Alexander ZAMBRANO. J'étends également mes remerciements à mes anciens professeurs de l'Université Distrital en Colombie : Antonio HERNANDEZ, José Luis HERRERA, Olga VARGAS, German RAMIREZ, Orlando RIAÑO et Gonzalo RICARDO, qui m'ont appris à voir le monde à travers des images, et pour m'avoir donné l'envie de poursuivre dans ce domaine. Je remercie aussi German CIFUENTES et Yull SALCEDO pour avoir soutenu ma candidature au master TGAE qui a déclenché cette aventure.

Je souhaite exprimer mes remerciements à trois belles femmes qui m'ont fait participer à leurs réseaux (familiaux et sociaux), et qui m'ont fait découvrir la France. Sans

elles mon séjour n'aurait pas été si sympathique ; nous avons été connectées par différents aspects. Les aléas de la thèse m'ont permit de partager avec Sophie PUJOL ; la passion pour les langues m'a liée à Martine FOURNIER ; et le risque et l'aventure m'ont permit de rencontrer Françoise PELISSON. Merci pour votre amitié, pour votre temps, pour votre soutien,…, pour les relectures et corrections de français. Je remercie toutes les familles et toutes les personnes qui m'ont accueillie chez elles ou ailleurs, pour un repas, un goûter, une promenade, un mariage, un café, une répétition de théâtre, une séance de squash, ou tout simplement pour une discussion ; ces moments ont nourri mon esprit et m'ont remplie de beaux souvenirs. Lucia, Maria et Andrea ont été mon contact avec mes racines pendant ce séjour, merci pour les agréables moments partagés : pour les repas, l'écoute, les voyages,…, pour avoir été là !. Paul, Patrik et Yoann merci pour votre soutien lors de mon séjour.

Finalement, je me réserve le plaisir d'exprimer un grand merci à ma famille : Gloria, Santiago, Laura, Lala et mon père, pour avoir été présents, sans être sur place, pour m'avoir encouragé à aller jusqu'au bout ; pour m'avoir soutenu ; pour m'avoir appelé dans les moments difficiles. A ma tante Java qui m'appelé de temps en temps. Je remercie aussi ma famille étendue : mes amis, pour m'avoir accompagnée et encouragée. Merci Andres Fdo, Angace, sr. Castillo, Cathy, Claus, Daniel A., Go, John, José, Liliana, Miller, Omitar, Oscar Guzman, Rocci, Sabry, Tamareto, Vasquez et Yaz.

Un immense merci à Javier qui m'a accompagné chaque jour durant cette thèse, qui m'a toujours soutenue inconditionnellement, qui a cru en moi quand moi je n'y croyais pas … pour cela et pour le reste, merci mon prince charmant.

Finalmente, me reservo el placer de expresar un gran agradecimiento a mi familia: Gloria, Santi, Laura, Lala y mi papá, por haber estado presentes, sin estar aquí, por motivarme a cumplir la meta, por haberme apoyado, por haberme llamado en los momentos difíciles. A la tía Java por llamarme de vez en cuando. Agradezco también a mi familia extendida: mis amigos, por haberme acompañado y apoyado.

Un inmenso agradecimiento para Javier, quien me acompañó todos los días durante esta tesis, quien me apoyo incondicionalmente, quien creyó en mí cuando ni yo misma creía…por eso y por todo lo demás, gracias mi príncipe azul.

Sommaire

Introduction

L'épidémiologie, selon l'Organisation Mondiale de la Santé (OMS, 1991), est considérée comme « l'étude de la répartition des problèmes relatifs à la santé de la communauté et à la maladie ainsi que leurs déterminants dans les populations humaines ». Afin de promouvoir la santé et réduire (voire prévenir ou combattre) la maladie, cette discipline scientifique s'appuie sur trois volets : recueillir, interpréter et utiliser l'information. En tant que discipline, l'épidémiologie s'est développée après la seconde Guerre Mondiale, même si l'approche épidémiologique a eu différents précurseurs : Hippocrate, John Graunt, Pierre Charles Alexandre Louis, William Farr, et John Snow, entre autres (CDC, 2006 ; Czernichow *et al.*, 2001 ; Merril, 2008 ; Ravaud, 2006). Hippocrate (460 avant notre ère) dans son œuvre « De aëre, aquis et locis » (*Des airs, des eaux et des lieux*, en français), suggère que certains facteurs comme l'environnement, l'alimentation et le comportement, en particulier, peuvent influencer le développement de la maladie. John Graunt (1620-1674), en prenant appui sur des registres paroissiaux de baptêmes et d'enterrements, a construit des tableaux chronologiques des naissances et des décès, et a également estimé la population à Londres[1], sous l'hypothèse d'un rapport constant entre le nombre de morts et de vivants, avançant ainsi la méthode du calcul du taux. Pierre Charles Alexandre Louis (1787-1872), intéressé par la mesure et la quantification des maladies, a développé la « méthode numérique » qui consiste en la standardisation du recueil et de l'analyse des données.

[1] John Graunt, 1665, « Natural and political observations mentioned in a following Index, and made upon the Bills of Mortality », Royal Society, Londres, 205p.

William Farr (1807-1883) est reconnu pour la classification des maladies en fonction de leur localisation anatomique. John Snow (1813-1858) a démontré le rôle primordial de l'eau, comme source d'infection et véhicule de transmission, dans l'épidémie de choléra qu'a connue Londres de 1849 à 1854. Ses travaux sont reconnus car ils illustrent la progression « classique » de la démarche de l'épidémiologie qui va de l'observation, en passant par l'élaboration d'hypothèses, le test de ces hypothèses, jusqu'aux applications, c'est-à-dire les interventions en santé.

Actuellement, l'épidémiologie est divisée en quatre approches (OMS, 1991) : descriptive, analytique, expérimentale et évaluative. L'épidémiologie descriptive étudie la distribution et la fréquence des maladies (nous y reviendrons plus loin). L'épidémiologie analytique recherche les causes ou les déterminants des problèmes de santé, ainsi que le lien entre une maladie donnée et certains facteurs potentiels. L'épidémiologie expérimentale utilise les essais cliniques afin d'évaluer les médicaments ou les pratiques médicales visant à lutter contre les maladies. L'évaluation épidémiologique vise à mesurer l'efficacité des actions de santé mises en œuvre pour améliorer l'état de la santé de la population. Les approches analytique et expérimentale sont fondées sur les études de cas-témoins, les études de cohortes (le terme « cohorte » désigne un groupe d'individus qui ont vécu un événement commun et qui sont suivis dans le cadre de l'étude) et les essais contrôlés.

L'épidémiologie descriptive, afin de comprendre et de résoudre les problèmes de santé publique, cherche à répondre aux questions suivantes : quel est l'événement de santé ? quelle est sa fréquence ? quelles sont ses manifestations, ses caractéristiques ? Pour ce faire, elle décrit, à l'aide de variables ou de caractéristiques : « qui » affecte-t-il ?, « où » se situe-t-il géographiquement ?, et « quand » apparaît-il ? Parmi les variables qui décrivent la question « qui ? » on trouve : l'âge, le sexe, le niveau d'éducation, la profession, le groupe ethnique, etc., des individus touchés. Concernant les variables reliées à la question « où ? », on trouve : le type d'agglomération dans laquelle les individus habitent ou travaillent ; la proximité de certaines sources de pollution ; la nature géologique des sols ; la couverture végétale etc. Pour la question « quand ? », cela consiste à préciser la date d'apparition des nouveaux cas de la maladie et à prévoir à quel moment les phénomènes seront les plus accentués (« pics »). Pour mesurer l'importance des maladies et des problèmes sanitaires, l'épidémiologie descriptive utilise deux

mesures : l'incidence et la prévalence. Les cas incidents constituent les nouveaux cas d'une maladie ou d'un problème de santé apparus pendant une période de temps définie, dans une communauté donnée. Les cas prévalents, eux, caractérisent le nombre total de cas affectés par une maladie ou un problème de santé à un moment donné. Ils comprennent aussi bien les cas ayant contracté la maladie avant le moment de la mesure, que les cas encore affectés à ce moment ; ils informent, d'une certaine façon, sur la durée de la maladie. Etant donné que la distribution des problèmes de santé n'est uniforme ni dans l'espace (la population humaine sur tout le Globe) ni dans le temps, les nombres de cas (incidents ou prévalents) ne permettent pas, à eux seuls, de comparer la fréquence d'une maladie entre groupes ou entre communautés, ni même de mettre en évidence une tendance. Pour ce faire, il est nécessaire de rapporter les occurrences des maladies aux tailles respectives des populations, exprimant ainsi l'information en termes de taux d'incidence ou de prévalence. Le nombre de cas (incidents ou prévalents) figure au numérateur, tandis que la population exposée figure au dénominateur. Le calcul des taux permet (entre autres) : d'avoir une définition précise et standardisée de la maladie ; d'apprécier les besoins sanitaires d'une communauté ; d'établir la vitesse d'apparition des nouveaux cas ; de calculer le nombre de cas attendus dans une communauté ; et de comparer deux populations ayant un nombre différent d'individus exposés au risque. Ainsi, le calcul des taux permet de définir les axes de recherche sous l'angle de l'étiologie des maladies, c'est-à-dire de la relation entre facteur de risque et apparition d'une maladie.

S'agissant de la population, celle-ci est dynamique au cours du temps : naissance, décès, immigration, émigration, évolution de la pyramide des âges (Bernard et Lapointe, 1987) ; de plus, selon la nature de l'étude épidémiologique, la population peut se classifier en : population fermée ou population ouverte (Bouyer *et al.*, 1995). Ces populations se caractérisent par les critères d'entrée (critères auxquels l'individu doit être soumis pour être considéré comme sujet de l'observation) et de sortie (critères à partir desquels l'individu n'est plus observé). Dans une population fermée, les critères d'entrée s'appliquent au début de l'étude, et les critères de sortie sont le décès ou la production de l'événement d'intérêt, aucun individu ne pouvant s'ajouter après la période d'admission. A l'inverse, dans une population ouverte, les critères d'entrée sont applicables à un moment quelconque dans la période d'observation ; de même que

l'admission à l'étude, le critère de sortie consiste en l'inaccomplissement d'un des critères d'entrée. L'information de la population est recueillie, en épidémiologie, à partir de sources directes (comme dans l'étude de cohortes) ; ou à partir de sources indirectes (comme des données démographiques, ou celles de l'état civil). Le type de source pris en compte définira la portée de l'évaluation de l'état de santé d'une communauté. Le recensement, ou dénombrement détaillé des habitants d'une région à un moment donné, constitue l'une des sources d'information de base de la démographie (Chesnais, 1990 ; Dumont, 2004). Il est utilisé, bien que présentant des limites et des défauts (nous reviendrons sur ces limites), afin d'obtenir la taille des populations ouvertes. Enfin, selon l'OMS (1991), la connaissance de la densité et de la distribution de la population d'une communauté est bien entendu importante pour la planification des services de santé, en particulier l'implantation de nouveaux dispensaires et de centres de santé ; pour l'évaluation de l'accessibilité et de la couverture de différents programmes sanitaires ; et pour l'évaluation des activités de promotion de la santé et de la lutte contre les maladies. L'évaluation de l'état de santé de la communauté fournit des éléments pour développer des politiques et des plans de santé publique qui, à leur tour, permettront d'améliorer l'état de santé de la population atteinte.

Le déroulement d'une campagne de recensement comprend plusieurs étapes, entre autres : l'ensemble des opérations d'organisation et de préparation de l'enquête ; la réalisation et le contrôle de la collecte des informations sur le terrain ; et l'exploitation des informations afin de connaître l'effectif total de la population (INSEE, 2005 ; Mignet, 1956). Chacune des étapes est constituée de différentes tâches et inclut de nombreux acteurs. En raison de la taille de la population, le temps de réalisation des campagnes de recensement peut varier entre plusieurs mois et plusieurs années, avec un coût budgétaire qui peut être considérable : cela limite la mise à jour des recensements. Cette difficulté ne concerne pas seulement les pays en voie de développement : en France par exemple, le recensement général prévu en 1997 a dû être repoussé à 1999 pour raisons financières (INSEE, 2005). En outre, la mobilité de la population (immigration, émigration, lieux temporaires de résidence, entre autres), la difficulté d'accès à l'intérieur des immeubles (due à l'absence des individus, à la topographie, à l'insécurité …) représente un facteur important dans la couverture de la population au moment du recensement. Cette couverture n'atteint pas 100%, et en général les individus exclus correspondent,

entre autres, à : des groupes minoritaires (ethniques ou autres), des personnes handicapées ou souffrant d'une maladie de longue durée, des personnes isolées (Raleigh, 1994). Etant donné la fréquence des recensements (environ tous les 10 ans), le Bureau de recensement procède à des estimations de la population dans les périodes intercensitaires. Ces estimations reposent sur certaines hypothèses de croissance de la population sans prendre en compte certaines variables spatiales. D'ailleurs, le dernier recensement « traditionnel » en France a eu lieu en 1999 (depuis 1801 sa fréquence était de tous les 7-9 ans - Czernichow *et al*., 2001) ; postérieurement, un nouveau mode de recensement à été mis en œuvre via des enquêtes de recensement réalisées avec des modalités différentes selon la taille de la population de la commune : 10 000 habitants ou plus (INSEE, 2005). En outre, une autre composante a été inclue dans le calcul de la population (INSEE, 2009) :

> *L'introduction d'un ajustement est destinée à assurer la cohérence entre, d'une part, la variation de la population de la France déduite des résultats de deux recensements et, d'autre part, les composantes de cette variation, le solde naturel et le solde migratoire, estimées par ailleurs. L'ajustement constitue alors une troisième composante, fictive, de la variation de population, qui permet de caler les estimations de population sur les résultats du recensement. L'ajustement traduit ainsi un défaut de comparabilité entre les chiffres issus de deux recensements. Il peut être lié à une évolution de la méthode de recensement mais également aux évolutions mêmes de la société.*

Actuellement, les informations du recensement, source de données démographiques officielle (Raleigh, 1994), sont devenus une référence primordiale pour les décideurs, tant publics que privés. Or, le recensement n'apporte pas obligatoirement des données valides sur la population en raison des difficultés, et dans la préparation, et dans l'exécution, comme dans la mise à jour (Briggs *et al.*, 2007 ; Hsu, 1971 ; Kraus *et al.*, 1974 ; Li et Weng, 2005 ; Liu *et al.*, 2006 ; Ogrosky, 1975 ; Lo 1977).

Un exemple illustre cette situation (Denis et Moriconi-Ebrard, 2009). Au Nigeria, des campagnes de recensement ont eu lieu avec une certaine régularité depuis les années 1950. Néanmoins, la comparaison du recensement de 1952 (avec 288 unités locales de plus de 5 000 habitants) avec celui de 1963 (avec 2 113 unités locales de plus de 5 000 habitants) met en évidence de fortes surestimations distribuées de façon inégale selon les régions et les villes. De plus, le recensement suivant, en 1973, a été officiellement annulé et aucun résultat n'a été publié. Dans les années 1980, aucun recensement n'a eu lieu, et les résultats de celui de 1991 a généré des polémiques en

raison du décalage entre les dénombrements officiels et les projections existantes à ce moment-là. A propos du recensement de 2006, on peut constater qu'il n'est disponible qu'à l'échelle locale ; et au moment de la publication (Denis et Moriconi-Ebrard, 2009) ces résultats n'étaient pas confirmés.

Plus globalement, un aperçu de l'état des lieux des recensements au niveau mondial est présenté dans les travaux de Moriconi-Ebrard (1991) :

> *… la couverture statistique du monde révèle de grandes inégalités suivant les continents : l'Amérique et l'Europe disposent d'une couverture statistique pratiquement exhaustive et de séries fréquentes. Par contre, la production statistique est déficiente dans certaines parties de l'Asie et de l'Afrique, où il faut se contenter de séries anciennes ou de données peu fiables. Un des faits les plus remarquables est l'abondance de données publiées sur les pays sud-américains qui, dans ce domaine, ne sont pas loin d'égaler les grands pays développés, et dépassent en tout cas une grande puissance comme l'URSS. Au contraire, certains pays passant pour des "puissances régionales", comme le Nigeria ou le Viêt-Nam se rangent ici aux côtés des pays les moins avancés.*

En définitive, les difficultés rencontrées autour des recensements ont stimulé la recherche d'alternatives pour estimer la population à l'aide d'autres sciences, notamment, la géographie. Ainsi, différentes recherches géographiques ayant comme but l'estimation de la population ont été menées, et ce avec différentes approches, différentes sources de données, et à différentes échelles. Une grande partie de ces estimations vise à établir le nombre d'habitants dans de petites zones à différentes échelles car les données de population, en général, agrègent (Tobler, 1979 ; Mennis, 2003 ; Hardin *et al.*, 2007) le nombre d'habitants en unités administratives, par exemple communes ou zones urbaines.

Dans ce cadre, cette présente recherche naît du besoin des épidémiologistes d'avoir accès aux informations sur une population exposée à un risque donné, et ce afin d'évaluer l'état de santé des communautés, notamment pour calculer les taux d'incidence des maladies (Viel et Tran, 2009). Ainsi, nos travaux se veulent en continuité avec ces travaux de recherche transdisciplinaires, tendant à rapprocher différents domaines. Dans notre cas, nous avons décidé de « croiser » la télédétection (issue des Sciences de la Terre, de l'Environnement et de l'Univers), avec l'épidémiologie (venant des Sciences de la Vie et de la Santé), et avec l'analyse spatiale (provenant de la Géographie, Sciences

Humaine et Sociale). Nous visons une estimation de la population de la ville de Besançon à l'aide de données de télédétection à très haute résolution spatiale (THRS).

Certes, l'approche géographique pour l'estimation d'une population ne constitue plus une avancée scientifique, différentes recherches ayant été menées depuis les années soixante-dix (voire antérieurement). Toutefois, le côté novateur de nos travaux réside dans le rapprochement de la télédétection et de l'épidémiologie dans une recherche géographique appliquée à la santé publique.

Ce travail tente d'apporter des réponses aux questions suivantes. Les données télédétection permettent-elles d'estimer la population, celle-ci étant considérée comme le dénominateur dans le calcul des taux d'incidence, dans un cadre épidémiologique ? Si oui, quel type de données télédétection se révèle le plus adéquat pour opérer cette estimation ? Quelle est la précision des taux d'incidence calculés avec la population estimée à partir des données télédétection ? Est-il possible d'utiliser la population estimée pour calculer les taux d'incidence d'une maladie quelconque ? La méthode proposée pour estimer la population est–elle généralisable ?

In fine, le principal objectif de ce travail est de contribuer à l'estimation de la taille des populations humaines (à partir des données THRS) à des fins épidémiologiques, et ce à une échelle spatiale infra-communale, correspondant donc à des zones de petite taille. De plus, nous visons : d'une part, à évaluer l'apport des données THRS par rapport aux données à haute résolution spatiale (Landsat), dans un même cadre urbain (Besançon) ; d'autre part, à vérifier ces estimations dans le calcul des taux d'incidence (cancers).

Etant donné les visées médicales -de santé publique- de prévision et d'action sanitaire, cette estimation doit être simple, automatique, reproductible et exportable. Cela va dans le sens de la Charte internationale pour la gestion des catastrophes[2], où une cartographie rapide des zones urbaines est cruciale, de nombreux prestataires de services spatiaux étant aujourd'hui disponibles pour faire face à toute demande (Gamba *et al.*, 2011).

Dans l'estimation de la population, cette étude utilisera pour sa réalisation uniquement les données THRS, parce qu'elles ne dépendent pas des contraintes des

[2] Charte relative à une coopération visant à l'utilisation coordonnée des moyens spatiaux en cas de situations de Catastrophe Naturelle ou Technologique Rév.3 (25/4/2000).
http://www.disasterscharter.org/charter (consulté le 17 février 2012)

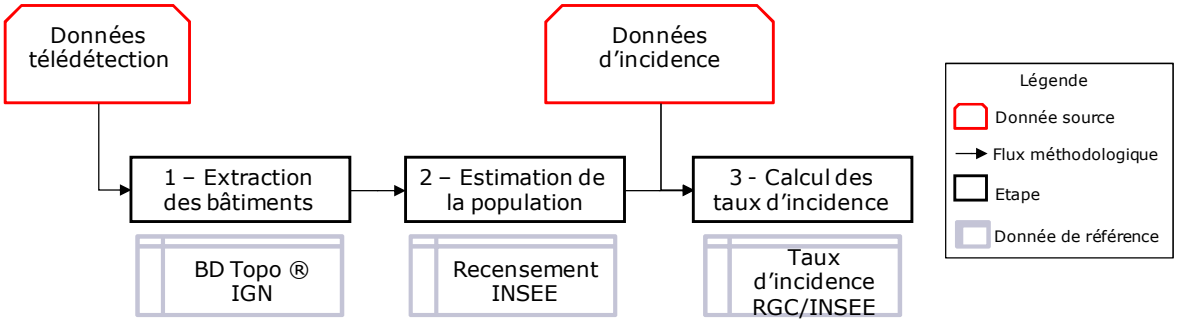
données de recensement, qu'elles peuvent être mises à jour plus régulièrement, et qu'elles sont d'une plus grande accessibilité aux applications d'usage civil. En outre, la diminution du temps d'acquisition des images satellite permet leur utilisation dans des conditions normales comme dans des situations d'urgence (Gamba *et al.*, 2011). Même si elles permettent peut-être d'obtenir de meilleurs résultats dans l'estimation de la population, nous n'utiliserons pas de données supplémentaires telles que celles liées à l'aménagement du territoire, comme le cadastre ou le plan d'occupation des sols (commerciale, industrielle, résidentielle, notamment). Cela aurait en effet considérablement limité l'exportabilité de nos résultats. En revanche, ces données seront utilisées comme « vérité terrain » pour valider notre approche. En outre, le calcul des taux d'incidence se limitera aux taux bruts, car on n'arrive avec des images qu'à une estimation de la population totale, sans pouvoir entrer dans le détail de sa distribution par âge et par sexe (qui permettrait de calculer des taux spécifiques et des taux standardisés). Par ailleurs, nous nous contenterons de mettre au point la méthodologie avec les données dans le cadre de Besançon, ville où nous disposons tout à la fois des données de télédétection (qui permettent de construire les modèles d'estimation de la population) et des données qui autorisent de contrôler la qualité des estimations. En effet, malgré les efforts entrepris pour collaborer avec d'autres laboratoires internationaux, en particulier en Colombie, nous n'avons pas obtenu l'ensemble des données nécessaires (sanitaires, censitaires et satellitaires) pour une même zone afin d'appliquer notre méthodologie.

Dans cette recherche, nous partons du fait que nous disposons *a priori* des données d'incidence, et donc nous ne nous intéresserons pas au recueil de ces données. Dans ce contexte, le choix de la ville de Besançon comme zone d'étude repose sur la disponibilité de données à la fois censitaires, médicales et d'imagerie spatiale. De plus, la diversité des types d'habitat (centre ancien, quartiers modernes, barres d'immeubles, zones pavillonnaires, entre autres), des densités de constructions (haute, moyenne et basse), et des couvertures du sol (végétation, urbaine, cours d'eau, *etc*.), fait de cette ville une zone d'étude qui garantit *a priori* une certaine exportabilité.

Afin d'atteindre nos objectifs nous proposons donc une méthodologie en trois étapes (fig. 0-1), fondée sur la corrélation existant entre la densité de population et la morphologie urbaine, reflétée sur les données télédétection (Harvey, 2002b ; Liu *et al.*,

2006 ; Viel et Tran, 2009 ; entre autres). Ainsi, la première étape consistera à extraire des bâtiments à partir des données télédétection ; ces bâtiments seront utilisés dans la deuxième étape pour modéliser la population ; à leur tour, ces populations serviront de dénominateur, lors de la dernière étape, pour calculer des taux d'incidence. Différentes approches seront utilisées dans les deux premières étapes, afin de couvrir un certain nombre de scénarios possibles. Pour l'extraction des bâtiments, les classifications par pixel et orienté-objets seront expérimentées, tant dirigées que non dirigées. Pour l'estimation de la population, des approches surface et volume seront employées. Quant au calcul des taux d'incidence, un cancer relativement fréquent (le cancer du sein), et un cancer plus rare (lymphome non-hodgkinien) seront considérés. Enfin, chaque étape sera validée avec des données de référence, et des analyses de sensibilité de la méthode seront effectuées.

Figure 0-1 Méthodologie générale



Notre exposé sera fortement lié à la méthodologie générale afin de maintenir l'aspect intégratif de cette thèse transdisciplinaire. Nous commencerons par faire le point sur la question (chapitre 1) ; suivront la zone d'étude et les données dont nous disposons pour mener cette recherche (chapitre 2). Ensuite, nous entrerons dans la première étape de la méthodologie générale, la télédétection et l'extraction des bâtiments (chapitre 3). Puis, nous aborderons l'estimation des populations par une approche dasymétrique -deuxième étape de la méthodologie générale- (chapitre 4). Enfin, nous envisagerons la dernière étape de notre méthodologie, l'épidémiologie, où les indicateurs épidémiologiques occuperont un rôle central (chapitre 5). Pour finir, les conclusions et les perspectives de cette recherche seront exposées.

Chapitre 1. Etat de la question

Nous aborderons l'état de la question à partir des trois lignes de force de cette recherche (télédétection, modélisation de la population, et épidémiologie), tout en se centrant sur notre fil conducteur : l'estimation de la population.

En premier lieu, différentes recherches géographiques menées sur la question de l'estimation de la population proprement dites seront exposées ; ensuite, nous présenterons les différents travaux réalisés en télédétection comme outil de construction des variables pertinentes dans l'estimation de la population ; enfin, nous considérerons l'utilisation de la télédétection dans le domaine de la santé publique, notamment en épidémiologie. Etant donné l'interdisciplinarité de cette recherche, nous ferons mention des aspects théoriques nécessaires au fur et à mesure de notre progression, afin de la rendre plus compréhensible.

1.1 Estimation de la population par la recherche géographique : données et méthodes

A son origine, le recensement était utilisé, entre autres, à des fins militaires ou économiques. Par exemple, selon Durand (2000) entre les XIX^{ème} et XVIII^{ème} siècles avant notre ère, au royaume de Mari (Syrie), « *(…) il existait des recensements nominaux où les propriétaires étaient fichés en fonction de leur statut militaire (…)* » ; et ce peut-être, afin de savoir combien d'hommes étaient disponibles pour défendre l'Empire. Du IV^{ème} au II^{ème} siècle avant notre ère en Chine, différents recensements ont été effectués (Mignet, 1956) « *(…) La connaissance de l'importance de la population était utile pour la détermination de la matière imposable, la satisfaction des besoins militaires et l'exécution*

des travaux publics (Grande Muraille (…) ». D'ailleurs, le premier dénombrement de la population chinoise s'est fait au XXIII^{ème} siècle avant notre ère. Il existe, également, la trace de dénombrements antiques dont on ne connait pas l'objectif, comme selon Birot (1958) le recensement des femmes du royaume de Mari effectué probablement sous le règne de Zimrilim, découvert dans la salle 5 du Palais de Mari. En France, le pouvoir royal a fait dénombrer les paroisses et les feux dès 1328 : le recensement est devenu systématique et régulier avec l'initiative de Napoléon 1er au XIX^{ème} siècle (Czernichow et *al.*, 2001).

Wu *et al.* (2008) proposent trois méthodes pour estimer la population (ou la distribution de la population), en fonction des données utilisées. Cependant, il faut noter qu'un certain nombre de recherches combinent celles-ci. La première utilise uniquement les données du recensement, et nous la nommerons « interpolation spatiale ». La deuxième utilise uniquement des variables physiques ou socio-économiques utiles pour inférer la population, nous la nommerons « modélisation statistique ». La dernière catégorie utilise ces deux séries de données : le recensement et les variables physiques ou socio-économiques utiles pour définir la population, ce que nous nommerons « méthode dasymétrique ».

1.1.1 L'interpolation spatiale

Le terme « interpolation spatiale » (IS) fait référence, globalement, à la question de l'estimation de la valeur d'une variable z à un point donné (x, y), étant donné les valeurs connues de la variable z d'un certain nombre de points autour du point à estimer. Comme telle, l'IS est inhérente aux courbes hypsométriques ou de niveaux (Goodchild et Lam, 1980) : la première carte de ce genre, produite par Pieter Bruinss en 1584, comprend les lignes isométriques de profondeur de la rivière Sparnee en Hollande (Maceachren, 1979 ; Imhof, 1982). Toutefois, la représentation de variables culturelles ou socio-économiques a trouvé sa place dans l'IS. Lalanne (1845) a ainsi explicité ce qui

pourrait être l'origine de la méthode utilisée pour produire les cartes isoplèthes[3] pour la population :

> *Supposons, en effet, que l'on partage le territoire d'un pays en un très grand nombre de parties suffisamment petites, telles que les fournirait la division par communes par l'étendue de la France ; qu'au centre de chacune de ces parties on élève une verticale proportionnelle à la population spécifique, ou, en d'autres termes, au nombre d'habitants par kilomètre carré, dans le territoire de la commune que l'on considère ; que l'on réunisse par une surface courbe continue les extrémités de toutes ces verticales, et qu'enfin on projette sur une carte, à une échelle convenable, et que l'on cote les courbes de niveau tracées sur cette surface et qui correspondent à des hauteurs verticales entières et équidistantes : on aura ainsi les lignes d'égale population spécifique, et on distinguera de suite la série des points où la population est de 30, 40, 50, ..., 100 habitants par kilomètre carré.*

Il faudra toutefois attendre 1857 pour que la première carte isoplèthe de la population soit produite et publiée par Ravn (Tobler, 1979).

On peut classifier l'IS en deux grands groupes (Lam, 1983): l'interpolation fondée sur des données de type « point » (de point à polygone) et l'interpolation de type « surface » (de polygone à polygone), chacun de ces groupes utilisant différentes méthodes. En ce qui concerne l'interpolation de données de type point, les méthodes peuvent être classifiées en : « exactes » ou « approximatives ». L'interpolation exacte repose sur l'hypothèse que les valeurs connues de la variable z sont fiables, et donc elles sont conservées lors de l'interpolation : les méthodes « distance pondérée », « krigeage », « interpolation par splines », et « différences finies » font partie de ce groupe. Dans le cas de l'interpolation approximative, les valeurs connues de la variable z présentent une certaine incertitude et pour cette raison la tendance globale de points et le lissage des erreurs sont pris en compte dans cette méthode, et donc les valeurs connues de la variable z ne sont plus conservées lors de l'interpolation : les méthodes « surfaces de tendance », « modèles de Fourier », « distance pondérée par moindres carrés », et « moindres carrés ajustés avec splines » font partie de ce groupe.

En ce qui concerne **l'interpolation de données de type « point »**, ces interpolations affectent la population d'une zone du recensement à un point donné de

[3] Ligne situant des zones de valeurs dont le tracé est établi par rapport à des points en nombre limité et de valeur déterminée ou estimée ; elle est utilisée pour mettre en évidence les variations de taille moyenne des grains, de leur sphéricité, etc. Source : Dictionnaire des sciences de la terre par M. Moureau et G. Brace. Editions TECHNIP, Paris, 2000.

cette zone (centroïde), puis les surfaces de populations, représentées et stockées par grilles, sont générées à partir de ces points (Bracken, 1991 ; Wu, 2008). En outre, ces grilles permettent de représenter les zones non peuplées avec une population de zéro habitant (Martin, 1989). La méthode proposée (Bracken et Martin, 1988 ; Martin, 1989) repose sur trois hypothèses : a) un centroïde définit une localisation avec une densité de population au-dessus de la moyenne de la zone, dont il est un point de synthèse ; b) un centroïde de la population est distribué dans les environs selon certaines fonctions décroissantes de distance, qui ont une extension finie ; et c) il existe des régions (dans cette approche de type point) pour représenter, a priori, les zones sans population, par exemple les zones d'eau. En Grande Bretagne, par exemple, les centroïdes représentant la population ont été distribués à travers les secteurs administratifs (districts) : ils ont été localisés, à la main en fonction des secteurs postaux, sur les maisons ou les pâtés de maisons à usage résidentielle même si elles étaient pavillonnaires. Ces centroïdes font partie du dossier du recensement de chaque district. Les centroïdes du recensement de 1981 ont été utilisés pour générer une grille avec des cellules de 200 mètres pour distribuer la population dans les 53 cantons qui constituent la Grande Bretagne, de même que trente autres variables socioculturelles (Bracken, 1991). Plus récemment, en 2005, Harris et Chen ont étudié l'effet de la modification des paramètres dans l'algorithme de distribution de la population, notamment la distance entre centroïdes, et la taille de la fenêtre d'interpolation, dans l'estimation de la densité de quatre villes en Grande Bretagne, à savoir : Bristol, Norwich, Peterborough et Swindon. Ils ont découvert que le paramètre qui influence le plus la surface de population modélisée était la taille de la fenêtre.

S'agissant des méthodes par **interpolation des données de type « surface »**, celles-ci sont classées selon la propriété « pycnophylactique » (conservation de la masse ou conservation du volume), proposée par Tobler (1979). L'hypothèse sur laquelle repose cette propriété est qu'il existe une fonction de densité nommée *Z(x,y),* laquelle possède un nombre fini de valeurs pour les positions *(x,y),* et qui est également non-négative. C'est-à-dire que la somme de toutes les valeurs des positions *(x,y)* doit être égale à la valeur de la densité *Z(x,y).* Avec cette propriété, Tobler a proposé une approche pour résoudre le problème, pratique et fréquent, de la conversion, ou de la compatibilité, des

données recueillies par divers organismes gouvernementaux ayant des limites géographiques distinctes pour la même partie du monde. En outre, Goodchild *et al.* (1993) ont modélisé le couplage de la population de l'Etat de Californie divisé en 58 cantons (divisions territoriales) avec 12 zones d'étude hydrologique qui suivent le contour des lignes de partage des eaux (fournies par le Département des Ressources en Eau de Californie, siglé CDWR en anglais). La méthode d'interpolation appliquée dans cette étude a consisté à pondérer directement les surfaces des cantons aux bassins, en supposant une densité de la population uniforme dans les cantons. Ces suppositions posent un problème majeur dans ces types d'interpolations, car les distributions homogènes apparaissent rarement dans le monde réel (Wu *et al.*, 2005). Cela entraîne des erreurs grossières dans les zones non-homogènes (Goodchild *et al.*, 1993). L'utilisation des réseaux irréguliers de triangles (siglé TIN en anglais) a été proposée dans l'interpolation des données de type surface et a été testée sur les données du recensement en Allemagne (Rase, 2001) : les TIN respectent la propriété de la conservation du volume.

1.1.2 La modélisation statistique

La modélisation statistique cherche à déduire/induire les relations entre la population et d'autres variables (physiques ou socio-économiques) afin d'estimer la population d'une zone (Wu *et al.*, 2005). Dans cette approche, la connaissance *a priori* de données de la population est nécessaire, bien que la population, en elle-même, ne fasse pas partie du modèle d'estimation : toutefois, elle permettra de juger ces estimations. Wu *et al.* (2008) classifient les variables pertinentes pour estimer la population en cinq groupes : « zones urbaines », « utilisation du sol », « unités de logement », « statistiques du pixel de l'image », et « variables autres ». Nous trouverons ces variables combinées dans un certain nombre de recherches. Dans l'exposé des différents travaux, nous regrouperons les variables « utilisation du sol » et « unités de logement » en un seul groupe, « unités de logement », car dans la modélisation de la population, elles partent du même principe.

En ce qui concerne **« les zones urbaines »** (zone/tache bâtie de l'agglomération), une des premières relations établies avec la population a été formulé par Nordbeck en

1965. Il a avancé le fait qu'utiliser la croissance allométrique pour estimer la population d'une agglomération au moyen de sa zone bâtie est pertinent. Ainsi, la « loi de croissance allométrique[4] » (loi de croissance relative - Huxley 1932, 1950 ; Gayon, 2000), découverte en Biologie, s'est introduite dans le domaine de la géographie. Cette loi a été adaptée, de sorte que : $A = a.P^b$, où *A* est la zone bâtie d'une agglomération (en hectares) et *P* sa population ; étant donné leurs dimensions, la constante *b* devrait être de 2/3 (Nordbeck, 1971). En utilisant les zones bâties des agglomérations en Suède et leurs recensements des années 1960 et 1965, Nordbeck a testé la loi de la croissance allométrique ; la corrélation entre *logA* et *logP* était très forte. Une autre approche de la modélisation statistique de cette loi a été proposée par Tobler (1969). Cette fois, le raisonnement repose sur le fait que les zones bâties d'une agglomération peuvent être représentées par un cercle de rayon (*r*), lequel doit être proportionnel à sa population en suivant la formule $r = a\ P^b$, où P est la population. En utilisant les coefficients a = 0,035 et b = 0,44 (trouvés de manière empirique) dont le rayon a été mesuré en kilomètres, différentes cartes de population ont été produites dans le Michigan (Etats-Unis), dans le sud de l'Allemagne, dans le centre de l'Espagne et de la Turquie. De même, la population de Dallas a été estimée, à travers le rayon de cette agglomération. Une autre approche de cette loi a été proposée par Lo et Welch (1977) qui ont résolu l'équation par la méthode des moindres carrés. Ainsi, en utilisant la série des recensements des années 1951, 1956, 1961 et 1966 et le contour des villes de Taiwan, ainsi que le recensement de 1953 et le contour de 124 villes de Chine Continentale, le coefficient *a* et le exposant *b* ont été calculés. De même, les coefficients de corrélation ont ainsi été obtenus : ils varient entre 0,88 et 0,90 pour les villes de Taïwan tandis que pour les villes de la Chine Continentale ils étaient égaux à 0,75. Ces modèles ont été comparés avec le modèle calculé à partir de 13 agglomérations dont la surface a été mesurée sur des images satellite prises entre 1972 et 1974, leur recensement datant de l'année 1970. Enfin, deux modèles ont été proposés afin d'estimer la population pour des villes chinoises ayant moins de 2,5 millions d'habitants : Le premier modèle, tiré des recensements des années 1951-1956 et des surfaces de 124 villes de la Chine Continentale (log P = 4,8733 + 0,7246 log A) ; et le second, tiré du

[4] Cette loi indique que la taille d'une partie du corps, *y*, est liée à celle de certains standard, *x* (soit tout le corps, le reste du corps sans « *y* » ou une partie de l'organisme sélectionné en tant que standard pour des raisons de commodité), selon la formule $y = bx^k$, où *b* et *k* sont constants.

recensement de l'année 1970 et des surfaces mesurées sur des images satellite de 13 villes de la Chine Continentale (log P = 5,3304 + 0,4137 log A ; r = 0,82). Ce dernier modèle était considéré comme moins fiable par Lo et Welch. Antérieurement, Ogrosky (s'inspirant du travail de Holtz *et al.*) avait inclut en 1975 trois variables additionnelles -en plus de la zone bâtie d'une agglomération- dans la recherche géographique de l'estimation de la population. Cela reposait sur l'hypothèse selon laquelle les grandes villes à proximité d'une agglomération exercent une certaine influence sur la population des autres agglomérations. Cette hypothèse pourrait être fondée sur une analogie avec la loi physique de l'attraction universelle (ou la loi universelle de gravitation[5]) découverte par Newton : dans l'analyse spatiale, à présent, l'application de cette analogie est appelée « modèle gravitaire » (Pumain et Saint-Julien, 2010a). Cette recherche a été développée dans 18 agglomérations de la région du Puget Sound (Etats Unis). Les trois variables additionnelles à la zone bâtie (*Ai*), étaient : a) le réseau de transport (*Li*) de l'agglomération vers les autres agglomérations (autoroutes, tramway et autres) ; b) la plus proche distance (*Dij*) entre l'agglomération et une grande agglomération (les agglomérations ont été classées en fonction de leur population et de leur taille) ; et c) la surface bâtie de la plus grande agglomération la plus proche (*Aj*). Les populations ont été tirées du recensement de 1970 et les autres variables ont été mesurées sur des photographies infrarouges. L'équation utilisée pour estimer la population (*Pi*) était $Pi = k \pm b1Ai \pm b2Li \pm b3Aj \pm b4Dij$, celle-ci a été résolue par régression linéaire multiple : le coefficient de corrélation r atteignait 0,964.

Quant au principe des **« unités de logement »**, cette méthode repose sur l'hypothèse selon laquelle la taille de la population (ou la densité de la population) est reliée de manière significative aux informations décrivant les sous-zones résidentielles en termes de type de logement et de densité (Green, 1956). Pour ce faire, des unités de logement (ou des catégories d'occupation du sol par unité de surface) sont multipliées par la valeur moyenne de la population y résidant (nombre de personnes ou densité de population par unité de surface). Différentes recherches ont été menées sous cette

[5] Cette loi énonce que deux corps ponctuels de masse m et M s'attirent avec une force (F) proportionnelle à leurs masses et inversement proportionnelle au carré de la distance (d) qui les sépare, selon la formule F = G*(m*M/d²).

hypothèse. Ainsi, Porter en 1956 (Li et Weng, 2005 ; Wu *et al.*, 2005) a estimé la population dans une région du Liberia : il a compté le nombre de huttes existantes à partir d'une photographie aérienne et a multiplié ce nombre par la moyenne des habitants par hutte, tirée de l'échantillonnage de l'enquête du terrain. Plus tard, en 1971, Hsu a estimé la population d'une région proche d'Atlanta (Etats-Unis). Son objectif était d'établir une méthodologie pour estimer la population dans les périodes intercensitaires, d'où les données-sources de son travail : le recensement de l'année 1950, des cartes topographiques et des photographies aériennes des années 1952 et 1968. L'unité de référence spatiale pour cette étude était la cellule, d'un quart de mile, qui constituait la grille de la zone d'étude. Chacune des cellules faisait partie d'une des trois catégories suivantes : « urbaine », « rurale » ou « semi-urbaine » ; à l'aide de ces catégories, le nombre moyen de personnes habitant dans des maisons a été tiré du recensement de 1950. En outre, le nombre de logements par cellule a été défini au moyen de la photo interprétation : l'erreur dans le dénombrement était inférieure à 5%. Enfin la densité de population par cellule a été calculée de la manière suivante : $\frac{\text{nb logements} * \text{hab par logement}}{4}$. Deux cartes de densité de population ont été produites (1952 et 1968), de même qu'une carte de croissance de population entre les années 1952 et 1968. En outre, Lo (1989) a, de son côté, essayé d'automatiser le processus de dénombrement des logements, en extrayant la densité de surfaces bâties (par cellules) sur des photographies en format *raster*. Cette expérience s'est faite à Rhode Island (Etats-Unis), et les erreurs atteintes dans cette extraction varient d'une part entre 2,5% et 6,94% (photographie noir et blanc) et d'autre part, entre -8,24% et -13,64% (photographie agrandie, prise avec une caméra de large format).

Par ailleurs, cette question a été abordée à travers l'utilisation du sol en détaillant le type d'occupation du sol. Aussi Kraus *et al.* (1974) ont-ils défini quatre catégories de type d'occupation du sol : « résidence unifamiliale » (*R1*), « résidence multifamiliale » (*Rm*), « terrain de caravaning utilisé comme résidence permanente » (*RTp*) et « zones avec d'autres utilisations » (commerciale ou industrielle en particulier). La densité de population moyenne de chacune des catégories d'occupation du sol a été établie à partir des échantillons tirés aléatoirement sur les données du recensement de 1970. Enfin la population de quatre villes en Californie a été calculée avec la formule $P = AR1 * DR1 +$

$ARm * DRm + ARTp * DTp$, où P était la population estimée ; AR, Arm et ARTp étaient les surfaces de chacune de catégories d'occupation du sol ; et *DR1*, *DRm* et *DRTp* correspondaient à la densité de population de chaque catégorie. L'erreur combinée des quatre villes dans l'estimation de la population était de 4,51% ; cependant, l'erreur totale pour chacune des villes atteignait 9%.

Plus récemment, Dong *et al.* (2010) ont modélisé la population, d'une région autour de Denton (Etats-Unis), au moyen des modèles de régression linéaire (par moindres carrées) ainsi qu'avec des modèles de régression géographiquement pondérée. Les variables indépendantes dans ces modèles étaient : les unités de logement, la surface des logements et le volume des logements. Ces variables ont été calculées pour les zones à usage résidentiel et commercial, tirées du plan d'occupation des sols (donnée source). Le choix des variables repose sur l'hypothèse selon laquelle l'utilisation des données de la hauteur pourrait constituer une plus-value importante dans les méthodes traditionnelles utilisées pour l'estimation de la population. Les valeurs du coefficient de corrélation des deux modèles ont montré une forte association entre la population et les variables indépendantes, même si la moyenne de l'erreur variait autour de 50%.

Dans les **« statistiques du pixel de l'image »**, l'éventail de variables étudié va de la réponse radiométrique de l'image originale jusqu'à l'application de formules complexes et du calcul des indices sur les canaux originaux de l'image. L'application de ces méthodes est un peu tardive par rapport aux autres approches exposées auparavant, car elles présupposent la disponibilité de données satellitaires (effective à partir de 1972).

En 1982, Iisaka et Hegedus se sont intéressés à la coïncidence entre la réponse radiométrique des images satellite et la densité de population. Deux images, une Landsat1 (prise de vue en 1972) et une Landsat3 (prise de vue en 1979), ont été couplées avec les recensements des années 1970 et 1975 respectivement, sur une zone du Kanto (Tokyo métropolitaine inclue). Pour ce faire, 88 grilles (taille 500m x 500m) ont été choisies comme échantillon, et un modèle de régression linéaire de la forme $\mathrm{P} = \mathrm{Ax1} + \mathrm{Bx2} + \mathrm{Cx3} + \mathrm{Dx4} + \mathrm{E}$ a été créé pour chaque image. P était la population dans chacune des grilles et x1, x2, x3 et x4 étaient les valeurs de réflectance de chacun des canaux de l'image Landsat MSS. Les résultats obtenus dans cette recherche ont montré une différence importante entre les estimations faites avec chaque couple de données, et

pour cette raison des corrections sur l'échantillon ont été introduites. Le coefficient de corrélation r était de 0,939 pour le modèle fait sur l'image de l'année 1972, tandis que r atteignait 0,899 pour le modèle de l'image de l'année 1979.

Lo (1992) a estimé la population à Kowloon (aire métropolitaine de Hong Kong) à travers des modèles de régression linéaire, fondés sur la radiance des trois canaux d'une image SPOT1 (prise de vue 1987). Pour ce faire, douze unités de planification tertiaire (siglé en anglais TUP, à savoir l'unité la plus petite pour les données du recensement à Hong Kong) ont été couplées avec les données du recensement de 1986. Ainsi, deux modèles de régression linéaire ont-ils été développés : les erreurs dans l'estimation de cette aire métropolitaine varient entre 735,10% et 854,93%. D'ailleurs, deux nouveaux modèles ont été développés, limitant la surface d'application aux zones résidentielles (haute et basse densité), voyant leurs erreurs varier de 647,75% à 718,17%.

Postérieurement, en 2002, Harvey, s'inspirant des travaux d'Iisaka et Hegedus (1982), ainsi que de ceux de Webster (1996), a estimé la densité de population des villes de Ballarat et Geelong, en Australie. Dans la ville de Ballarat, 138 districts ont été pris comme échantillon, tandis que dans la seconde ville, l'échantillon se composait de 225 districts. Pour cette étude, la variable dépendante était la densité de population ; la distribution normale a été choisie comme modèle de probabilité ; et les modèles des moindres carrés ordinaires avec des erreurs normales ont été utilisés. La validation des modèles de régression s'est faite de façon externe, en utilisant la ville de Geelong. Cette recherche a été développée en trois étapes. Dans la première étape, 42 variables explicatives ont été inclues : les moyennes de chacun des canaux d'images TM (*bi*), le carré de chacune des moyennes des canaux (*si*), 15 produits croisés de chacune des moyennes de canaux par $pij = bi * bj$ 15 autres couples du ratio entre canaux $\left(\frac{\mathrm{bi}}{\mathrm{bj}}\right)$et 15 couples du ratio différence-somme entre canaux $\left(\frac{\mathrm{bi}-\mathrm{bj}}{\mathrm{bi}+\mathrm{bj}}\right)$. Cinq modèles de régression ont été choisis : le premier n'incluait que les canaux originaux (R^2 = 0,5), afin de comparer les résultats avec les travaux d'Iisaka et Hegedus (1982) ; les autres modèles intégraient d'autres variables ; le coefficient de corrélation R^2 atteint 0,7. La transformation de la variable dépendante (racine carrée ou logarithme) a amélioré les valeurs de R^2, qui atteignent les valeurs de 0,8 et 0,9. Dans la deuxième étape, reprenant les suggestions des travaux de Forster, Barnsley et Webster, les modèles ont été fondés sur les mesures

de variabilité spatiale : la variance (v), l'écart-type (s) et le coefficient de variation ($c = \frac{\text{écart−type}}{\text{moyenne}}$). Quatre modèles ont été retenus et la valeur de R^2 était de 0,751. Dans la dernière étape, des modèles fondés sur les transformations spectrales ont été calculés. Pour chacun des échantillons, 80 variables ont été calculées ; parmi elles, 14 correspondaient à des transformations spectrales telles que : normalisation, proportion, proportion différence-somme, tonalité en coordonnées cylindriques et rectangulaires. Le reste des variables correspondait aux mêmes types de variables calculées dans les étapes précédentes : la valeur de R^2 atteignait 0,77. Les transformations -logarithmique et racine carrée- augmentaient la corrélation entre les données télédétection et les chiffres de référence de terrain de l'ordre de 0,75 jusqu'aux alentours de 0,92.

Une étude a été réalisée par Li et Weng (2005) à Indianapolis (Etats Unis), avec un échantillon de 658 groupes de pâtés des maisons (*block groups* en anglais), sur lequel 162 groupes (25% du total) ont été utilisés pour calculer les modèles de régression. Six groupes différents de variables issues de la télédétection ont été calculés : la radiance de chaque canal d'une image Landsat ETM+ ; l'analyse en composantes principales calculées sur les canaux originaux ; les indices de végétation ; l'analyse de mixture spectrale ; la température ; et les indices de texture. Différents modèles ont été testés utilisant ces variables (toutes ou en partie) ; de plus, d'autres tests ont été faits en divisant la densité de population en trois sous-groupes. Le coefficient de corrélation R^2 a été calculé pour chacun des modèles : sa valeur varie entre 0,130 (modèle avec toutes les variables et densité base) et 0,863 (modèle avec toutes les variables et densité moyenne). Enfin trois modèles (ceux fondés sur les sous-groupes de la densité) ont été retenus pour estimer la population de la ville : l'erreur relative était de 3,2%.

Plus récemment, Liu *et al.* (2006) ont étudié, dans le comté de Santa Barbara (Etats-Unis) sur un échantillon de 1578 pâtés de maisons, la corrélation entre la densité de la population et la texture d'une image au moyen de la régression linéaire. Afin de décrire la texture de l'image, trois méthodes ont été testées et comparées. La première méthode repose sur les six descripteurs de texture, fondés sur la matrice de cooccurrence des niveaux de gris (siglé GLCM en anglais). La seconde méthode décrit la texture en utilisant la semi-variance, laquelle a été construite sur les variations de distance entre 1 et 20 pixels pour chacune de différentes zones urbaines homogènes définies *a priori* pour

les auteurs (siglé HUP en anglais). La dernière méthode réalise différentes mesures spatiales sur chacune des HUP en relation avec : la végétation, la surface bâtie, et le reste de la couverture. Les coefficients de corrélation entre la densité de la population et la texture d'une image s'élèvent à 0,45 par la méthode GLCM, à 0,20 par la méthode de la semi-variance, et à 0,55 par la méthode de mesure spatiale.

Abordons maintenant **« les autres variables »** socio-économiques ou physiques pertinentes dans l'estimation de la population. Green (1956), dans son étude, a découvert une corrélation forte entre la morphologie urbaine et les variables socio-économiques. Cette étude repose sur l'hypothèse selon laquelle l'information décrivant les sous-zones résidentielles en termes de types de logement et de densité est sensiblement liée à la taille de la population et à la densité de population, cette information ayant une valeur prédictive quant à la structure socio-économique de ces sous-zones. Cette hypothèse a été confirmée par des études de corrélation dans le recensement de six villes des Etats-Unis couplées aux unités de logement photo-interprétées. Les résultats obtenus ont mis en évidence que la prévalence du logement unifamilial était liée au type de propriété, au revenu et au statut professionnel, tandis que la densité de logement unitaire élevée était associée aux groupes avec faible revenu. Aussi l'auteur a-t-il proposé, parmi les clés de photo-interprétation des images, l'identification de trottoirs, de sentiers, de chemins de randonnée, de garages, d'aires de stationnement, entre autres. Par ailleurs, le projet « LandScan Global Population » (Dobson *et al.* 2000) a produit, en 1998, une base de données pour le monde entier (1kmx1km de résolution spatiale) afin d'estimer les populations à risque, et ainsi de pouvoir répondre à une situation d'urgence découlant d'une catastrophe. Cette base de données a intégré différentes séries de données saisies comme variables dans la modélisation : les routes (le réseau routier) ; la pente du terrain ; l'occupation du sol ; la lumière nocturne ; et le recensement. La sélection des données repose sur les hypothèses suivantes : la densité de routes était en corrélation directe avec la densité de population ; la pente du terrain était inversement proportionnelle à la densité de la population ; l'occupation du sol était un indicateur de la densité de la population ; et la lumière nocturne modélisait la distribution de la population. Concernant la variable « réseau routier », Veregin et Tobler (1997) ont démontré l'existence d'une relation allométrique entre le nombre de segments des routes dans une « zone urbaine »

et la population habitant dans cette zone. Cette expérience a été testée sur 240 zones classifiées comme « urbaines » dans le recensement de 1980 aux Etats-Unis : le coefficient de corrélation R^2 obtenu dans cette modélisation était de 0,85. Quant à la variable « lumière nocturne », elle a donné lieu à différentes recherches. Sutton (1997) a utilisé les images satellite en prise de vue nocturne, fournies par le programme de défense des satellites météorologiques -capteur OLS (siglé en anglais DMSP OLS), dans le but de les corréler avec la taille de la population. Cette recherche a abouti à un modèle permettant d'estimer la densité de la population à partir des taches urbaines éclairées tirées des images DMSP :

$$Ln(\text{population de la tache}) = 3{,}353 + 1{,}359 * Ln(\text{Surface tache}).$$

La zone d'étude de cette recherche englobait tout le territoire des Etats-Unis ; les données du recensement de l'année 1990 ont permis la validation des résultats de cette étude, montrant une forte corrélation entre les taches urbaines éclairées et la population. Plus récemment, en 2001, Lo s'est également servi de ce type de données (DMSP) prises au-dessus de la Chine (mars 1996 et janvier 1997). Dans sa recherche, les modèles de population (allométrique et par régression linéaire) se sont faits à trois échelles spatiales différentes : province, région et ville. Dans la modélisation proposée par l'auteur, les taches éclairées (inférant les « surfaces bâties ») ont été catégorisées selon l'intensité de la lumière (en six groupes) dans le but de simuler le volume de chaque tache urbaine. Ses résultats ont montré une corrélation forte entre la densité de la population aux échelles spatiales de région et de ville (en utilisant le modèle allométrique) et la surface (de la tache éclairée) et le volume (de la tache éclairée) comme variables indépendantes.

Plus récemment, Lu *et al.* (2010) ont intégré la hauteur des bâtiments comme variable dans la modélisation statistique de la population à Denver (Etats-Unis). Dans sa recherche 72 IRIS[6] (*blocks* en anglais) ont servi pour calibrer les modèles, tandis que 22 IRIS (localisés dans une autre partie de la région) ont été utilisés pour leur validation. Deux modèles de régression linéaire ont été testés, l'un fondé sur la surface des bâtiments et l'autre fondé sur leur volume. Les valeurs des coefficients de corrélation obtenus varient : d'une part entre 0,85 et 0,94 pour l'approche fondée sur la surface ; d'autre part entre 0,92 et 0,93 pour l'approche fondée sur le volume.

[6] Ilots Regroupés pour l'Information Statistique.

1.1.3 La méthode dasymétrique

La méthode « dasymétrique » désagrège spatialement les données en zones plus détaillées pour les analyser, en utilisant des données auxiliaires (*ancillary data,* en anglais). Selon Robinson (1955), la période entre 1835-1855 pourrait être considérée comme « l'âge d'or» du développement de la cartographie géographique : à cette époque, au cours de ces vingt années, presque toutes les techniques connues aujourd'hui pour représenter les effectifs de la population, leur distribution, leur densité et leurs mouvements semblent avoir vu le jour. De même, il attribue à Henry Drury Harness (1837) la réalisation de la première carte « dasymétrique » correspondant à une carte de densité de la population en Irlande. Cette carte fait partie des trois cartes produites dans le second rapport présenté aux commissaires nommés pour examiner et recommander un système général des chemins de fer pour l'Irlande, deux d'entre elles concernant directement la population de l'Irlande, et la troisième représentant la circulation des marchandises. Cependant, Scrope (1833) a produit, dans son livre *« Principes de politique économique »,* une carte de densité de la population dans le monde entier classifiant les pays en trois catégories : entièrement peuplée avec plus de 200 habitants par mille carré, sous-peuplée entre 10 et 200 habitants par mille carré, et non peuplée avec moins de 10 habitants par mille carré. Cette carte selon Friendly et Dennis (2001) est le point de repère du début de la méthode « dasymétrique ». Toutefois, le début de l'estimation de la population par la recherche géographique au moyen de cette méthode « dasymétrique » a été marqué par l'apparition de la carte de distribution de la population de Cape Cod (Massachussetts, Etat-Unis), produite par Wright en 1936, dont la désagrégation s'est faite en fonction de l'utilisation du sol.

La désagrégation des données (en général), et notamment des données de population (dans notre travail), s'est faite à l'aide d'une variable quantitative donnée (autrement dit « données auxiliaires »), qui est d'une part agrégée par des unités géographiques (Maantay *et al.*, 2007), et d'autre part reliée à la distribution de la population (Wu *et al.*, 2008). La gamme des variables utilisées varie entre : le découpage administratif, le découpage électoral (Flowerdew and Green, 1991), le cadastre (Maantay *et al.*, 2007), le réseau routier (Xie, 1995), le code postal (Langford *et al.*, 2008), l'occupation du sol, ou quelques autres variables socio-économiques ou physiques

pertinentes. Dans la partie présente de cette thèse -l'état de la question- nous nous intéresserons aux seules variables physiques ou socio-économiques pertinentes dans l'estimation de la population, issues des données de télédétection. Aussi ces variables seront-elles classées en trois grands groupes : « occupation/utilisation du sol » (*land cover/land use* en anglais), « statistiques du pixel de l'image », et « variables autres ». De même que dans la catégorie « modélisation statistique », il est nécessaire d'établir, dans cette catégorie « méthode dasymétrique », une relation mathématique entre la population (c'est-à-dire le « recensement ») et la variable avec laquelle la désagrégation sera faite.

Au sujet de **« l'occupation/l'utilisation du sol »**, cette variable est la plus utilisée pour désagréger l'information spatiale du recensement (Wu *et al.* 2008). Parmi les recherches menées, nous trouvons les travaux de Menis (2003) qui a utilisé l'information de l'occupation du sol afin de distribuer la population dans cinq régions du sud-est de la Pennsylvanie (Etats-Unis). Sa recherche a été motivée par la grande variation entre zones urbaines et rurales, trouvée dans les cartes de population produites sur ces régions, où la taille des IRIS (*block groups* en anglais) et la densité de population ont été significativement différentes. De cette manière, trois catégories différentes de densité d'urbanisation ont été proposées : densité urbaine haute ; densité urbaine basse ; et non urbaine. La densité de la population de chaque IRIS a été redistribuée sur une grille de 100m*100m, avec la méthode d'interpolation de type « surface » conservant la propriété « pycnophylactique ». Chaque cellule (de la grille) a pris en compte : d'une part la différence relative de la densité de population de chacune des trois catégories d'urbanisation ; et d'autre part, le pourcentage occupé par chacune des trois catégories d'urbanisation dans la surface totale de chaque IRIS. Cette différence relative de densité a été établie via un échantillonnage de l'occupation du sol couplé avec la densité de population. Postérieurement, Langford (2006) a utilisé cette même méthode dasymétrique, reposant également sur trois catégories (rurale, suburbaine et dense), pour distribuer la population de la région du Leicestershire (Angleterre). Les données de population correspondent au recensement de 1991, spécifiquement au niveau du canton (*ward* en anglais) composé de 187 zones. Son travail compare différentes méthodes pour estimer la population : pondération par surfaces, régression et dasymétrique. Les

résultats de cette méthode (divisée en trois classes) ont été peu concluants bien qu'ils aient été supérieurs à ceux des autres méthodes.

Plus récemment, Viel et Tran (2009) ont estimé la population par IRIS sur la ville de Besançon (France). Leur travail s'est fondé sur le lien existant entre la morphologie urbaine et la densité humaine. Aussi ont-ils produit deux modèles de désagrégation spatiale de la population reposant sur l'occupation du sol au moyen de surface bâtie/non bâtie. La population a été estimée dans un cadre épidémiologique, comme une alternative pour trouver un dénominateur dans la définition de l'état de santé de la communauté de Besançon. La pertinence des populations estimées a été vérifiée avec le calcul du taux d'incidence des cancers du sein et du lymphome-non hodgkinien : les coefficients de corrélation interclasses varient entre 0,71 et 0,84.

S'agissant des **« statistiques du pixel de l'image »,** Harvey, (2002b) a mené une autre recherche sur les villes qu'il avait déjà étudiées (Ballarat et Geelong) en Australie. Dans sa recherche, la population de chaque district a été distribuée entre les pixels constituant les zones résidentielles/non résidentielles (classifiées précédemment), avec la formule $Pi = \frac{P}{n}; i = 1, \dots, n$ où *Pi* est la population initiale assignée au pixel *i*, *P* la population du district et *n* le nombre de pixels (résidentiels) dans le district. Cette distribution repose sur l'hypothèse d'une distribution uniforme, même si la population initiale (Pi) a été ajustée en prenant compte les différences spectrales de chaque pixel (résidentiel). L'ajustement s'est fait de manière itérative, et en supposant une relation linéaire entre la population et les canaux de l'image. Une fois les populations de référence calculées, différents modèles de régression ont été testés, utilisant la réponse radiométrique des pixels comme variable indépendante. Afin de mettre en évidence les avantages d'avoir le pixel comme unité spatiale dans l'estimation de la population, les méthodes fondées sur données agrégées (Harvey, 2002a) ont été comparées avec les méthodes fondées sur les données du pixel. Les coefficients de corrélation R^2 atteints avec les modèles de régression fondés sur les données du pixel varient entre 0,73 et 0,82, tandis que les coefficients R^2 fondés sur les données agrégées varient entre 0,45 et 0,84.

Quant aux « **variables autres** », utilisées afin de désagréger la population, Brigss *et al.* (2007) ont utilisé la lumière nocturne en complément de l'occupation du sol. Ces variables ont été couplées avec les données du recensement de l'Union Européenne : des modèles de régression ont été utilisés pour établir la relation entre le recensement, l'occupation du sol et la lumière nocturne.

Postérieurement, Wu *et al.* (2008) ont intégré la hauteur dans un modèle déterministe pour estimer la population dans une région de Texas (Etats-Unis). Dans cette recherche, l'estimation de la population s'est faite au moyen du volume des bâtiments (construit avec le contour des bâtiments et leur hauteur) et de l'occupation du sol. Ainsi, la population (*Pop*) a été calculée avec la formule : $Pop = \frac{\mathrm{Bdv}}{\mathrm{HuSpace}} * \mathrm{Occrate} * \mathrm{HdSize}$, où *Bdv* est le volume du bâtiment, *HuSpace* l'espace moyenne par unité de logement, *Occrate* la taxe d'occupation du logement, et *HdSize* la taille moyenne des ménages. Les trois variables *HuSpace, Occrate,* et *HdSize* ont été dérivées du recensement. L'évaluation de leur modèle de désagrégation a été effectuée à l'aide de sous-unités simulées : l'erreur d'estimation a été inversement proportionnelle à la taille des sous-unités.

Tableau 1-1 : Résumé des méthodes pour estimer la population par la recherche géographique

Estimation de la population par la recherche géographique : Résumé			
Méthode	« interpolation spatiale »	« modélisation statistique »	« méthode dasymétrique »
Données d'entrée	Recensement	Variables physiques ou socio-économiques utiles pour inférer la population	Recensement et variables physiques ou socio-économiques utiles pour définir la population
Description	Estime la population d'une surface donnée à partir des données connues du recensement	S'intéresse à déduire/induire les relations entre la population et d'autres variables	Désagrège spatialement le recensement en zones plus détaillées pour les analyser, en utilisant d'autres variables

…Suite			
Approches	Interpolation des données de type point Interpolation des données de type surface	Zones urbaines Unités de logement Statistiques du pixel de l'image Autres variables	Occupation/Utilisation du sol Statistiques du pixel de l'image Autres variables
Modèle mathématique	Interpolation spatiale	Allométrie Régression	Interpolation spatiale type surface Régression

1.2 La télédétection comme outil de construction des variables pertinentes dans l'estimation de la population

Tout au long de son Histoire, l'Homme a cherché le moyen de se situer dans l'espace ; de représenter son environnement, son entourage ; de connaître et de comprendre son territoire. Les cartes lui ont permis de représenter graphiquement sa relation avec le territoire. A l'origine, les cartes représentaient la perception de sa réalité : par exemple, les habitants du Nil gravaient la description de leurs terres sur tablettes de terre cuite, afin de les distribuer correctement suite aux crues de chaque année (Hutorowicz et Adler, 1911). Plus récemment, l'Homme s'est orienté vers la capture de cette information en découvrant dans la photographie le moyen de le faire : la photographie permet donc de fixer, voire rendre permanente, une image de survol d'une surface. En 1855, le photographe Gaspard-Félix Tournachon, dit Nadar, prend la première prise de vue en hauteur sur le toit du bâtiment de son atelier, marquant le début de la « photographie aérienne ». Trois ans plus tard, il récidive avec une prise de vue aérienne du « Petit Bicêtre (sud de Paris) » à partir d'un ballon qui sert de plateforme. Par ailleurs, il dépose une demande de brevet d'invention « pour un nouveau système de photographie aérostatique » (Carbonnell, 1968) de son procédé pour prendre des vues aériennes en ballon. Au XIX^{ème} siècle, l'utilisation de la photographie est intégrée à différents champs disciplinaires. Cependant notre étude ne s'attachera qu'au seul champ de la Géographie, pour laquelle le développement des appareils photographiques pour prises de vue

aériennes s'est fait en synchronisme avec le développement de l'industrie aéronautique (Diete, 1988). Après la seconde Guerre Mondiale, où la photographie a été très utilisée, la mise au point de la couverture de « photographies aériennes[7] » systématiques s'est faite ; ce qui, couplé avec la « stéréoscopie[8] », fournit à la cartographie une autre source d'informations pour représenter le territoire. Postérieurement et avec l'avènement de l'exploration spatiale, les Etats-Unis lancent en 1972 le premier satellite d'observation de la Terre « ERTS-1 » pour utilisation civile, avec un capteur MSS (module de balayage multi-spectral) de quatre canaux à résolution de 57mx79m. « ERTS-1 » -ensuite renommé « Landsat 1 »- fait partie du programme spatial Landsat conçu pour l'étude des ressources terrestres, notamment la sylviculture et les applications géologiques[9]. En 1986 le lancement du satellite « SPOT 1 », premier satellite du programme d'observation de la Terre décidé par les gouvernements belge, suédois et français[10], marque une nouvelle ère dans la télédétection, grâce à sa résolution de 10m dans la bande panchromatique et de 20m en trois canaux, ainsi qu'à sa capacité stéréoscopique sur le capteur HRV (Haute Résolution Visible). Postérieurement, en 1999, une nouvelle période dans la télédétection est marqué par le lancement d'« IKONOS » le premier satellite commercial de très haute résolution spatiale (THRS) avec une résolution de 1m dans la bande panchromatique et de 4m en quatre canaux[11]. Plus récemment, en février 2000, une mission topographique (*Shuttle Radar Topography Mission* (SRTM)[12], en anglais) a été lancée depuis le bord d'une navette spatiale équipée d'un radar (*RAdio Detection And Ranging*, en anglais) d'altimétrie. Le SRTM a obtenu des données altimétriques sur environ 80% des surfaces terrestres du Globe (Farr et Kobrick, 2000). Dès 2004, la large diffusion d'images satellite par le biais de technologies telles que

[7] L'expression « photographie aérienne » désigne toute vue prise depuis un engin se trouvant dans l'atmosphère, aéronef, avion ou fusée (Tricart et al. 1970). Les photographies aériennes comportent l'image du terrain et d'autres indications annexes. Les photographies verticales sont les plus intéressantes pour la cartographie et l'interprétation. Elles permettent une lecture facile des détails ainsi que des mesures (distances, dimensions, hauteurs) Elles offrent la possibilité de restituer des lignes ou des points. En outre, elles peuvent être lues en vision stéréoscopique, donnant ainsi une bonne appréciation du relief (Bakis, 1978).

[8] Procédé qui permet d'obtenir la sensation du relief à partir de deux images stéréoscopiques d'un objet, prises de deux points de vue différents. (CILF, 1997)

[9] http://landsat.usgs.gov/about_mission_history.php (consulté le 8 septembre 2011)

[10] http://www.cnes.fr/web/CNES-fr/1778-fin-de-vie-de-spot-1.php (consulté le 8 septembre 2011)

[11] http://www.geoeye.com/CorpSite/products-and-services/imagery-sources/Default.aspx (consulté le 8 septembre 2011)

[12] http://www2.jpl.nasa.gov/srtm/ (consulté le 14 janvier 2012)

Google Earth ou Google Maps a permis à l'Homme non seulement d'envisager de nouvelles applications dans des domaines non strictement géographiques, jusque-là peu explorés, mais aussi de continuer avec la même démarche de se situer dans l'espace : de représenter son environnement, son entourage ; de connaître et de comprendre son territoire.

Au long de l'évolution et du développement de la photographie (aérienne, spatiale...), le terme « télédétection » a été introduit en 1971 pour remplacer « détection à distance », et postérieurement (en 1973), le terme « télédétection » a été reconnu officiellement (Provencher et Dubois, 2007). Sa définition selon *« Le manuel terminologique didactique de télédétection et photogrammétrie »* est la suivante :

> *Ensemble des connaissances et techniques utilisées pour déterminer des caractéristiques physiques et biologiques d'objets par des mesures effectuées à distance, sans contact matériel avec ceux-ci.*

Néanmoins, une définition, plus précise et largement acceptée, de la télédétection est celle donnée par le Centre Canadien de Télédétection[13],

> *La télédétection est la technique qui, par l'acquisition d'images, permet d'obtenir de l'information sur la surface de la Terre sans contact direct avec celle-ci. La télédétection englobe tout le processus qui consiste à capter et à enregistrer l'énergie d'un rayonnement électromagnétique émis ou réfléchi, à traiter et à analyser l'information, pour ensuite mettre en application cette information.*

Pour faire le point sur les travaux liés à la recherche géographique pour l'estimation de la population, notamment par exploitation des données de télédétection (photographie aériennes, image satellite ou autre), nous retiendrons trois grands groupes : le premier correspond à l'interprétation visuelle ; le deuxième s'intéresse au traitement d'image numérique ; et le dernier groupe s'attache à la modélisation des données de hauteur des objets géographiques localisés à la surface du sol.

1.2.1 L'interprétation visuelle

L'interprétation visuelle correspond à la « *méthode de lecture et d'analyse des images en général qui fait essentiellement appel à l'œil et aux capacités psycho-visuelles d'une personne (CILF, 1997)* ». Cette méthode a été largement utilisée pendant l'essor de

[13] http://www.cct.rncan.gc.ca/resource/tutor/fundam/chapter1/01_f.php (consulté le 21 septembre 2011)

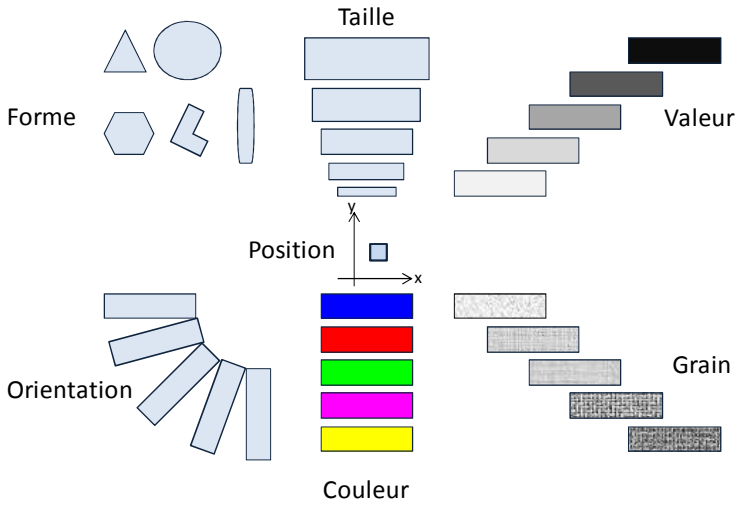
la photographie aérienne, et elle est encore utilisée avec des images de satellite, notamment celles à très haute résolution spatiale (THRS). D'un point de vue sémiologique, la photographie aérienne est une image figurative qui fait appel à l'œil comme système de perception et sur laquelle l'Homme attribue des signes polysémiques. La lecture d'une image se situe entre le signe et sa signification (Bertin, 2005). Gagnon (1974) explicite les étapes nécessaires dans le processus mental complet pour la photo-interprétation, processus décomposé en cinq étapes : 1) détection, 2) identification, 3) analyse, 4) déduction et 5) classification. Gagnon définit ces étapes ainsi :

> *La détection consiste à distinguer un objet ou un élément parmi ceux qui l'environnent, l'identification consiste à identifier un ou plusieurs objets ou éléments clairement visibles en raison de leur analogie avec des choses connues. L'analyse c'est le groupement en zones d'objets ou d'éléments de même nature. La déduction se situe à un niveau plus élevé d'abstraction, et se base souvent sur des indices convergents. Elle est l'obtention d'une information non directement observable sur la photo, obtenue grâce à d'autres observations faites sur la photo et à des connaissances provenant d'autres sources. … La déduction doit s'appuyer sur des vérifications faites sur le terrain. La classification est la description précise et systématique des surfaces délimitées par l'analyse.*

Plus récemment, ces étapes ont été regroupées en trois « stades d'interprétation » (Puissant, 2003) : 1) la détection, 2) l'identification et 3) l'analyse, où cette dernière comprend la déduction et la classification.

Ainsi, si l'on part d'un système graphique de signes monosémiques pour analyser les taches qui sont représentées sur une image, on dispose de huit variables visuelles (fig. 1-1) (Bertin, 2005) : 1) la dimension x, 2) la dimension y, ces deux composant les dimensions du plan, 3) la taille, 4) la valeur, 5) le grain, 6) la couleur, 7) l'orientation, et 8) la forme.

Figure 1-1 : Variables visuelles. Source *Bertin, 2005*



Dans ce système, la «taille rend compte de la variation de la surface représentée, tandis que la valeur fait référence à la variation des gris dans une échelle du noir au blanc. Le grain correspond à un semis régulier d'un motif quelconque contenu dans une surface unitaire. Cette variable est liée à la variation de taille et de valeur. La couleur est la perception visuelle des différentes longueurs d'onde qui constituent la lumière visible. Le ton est défini par deux paramètres : la couleur et la valeur. L'orientation est l'angle formé entre la tache et l'axe vertical. La forme est définie par le contour de la tache. Ces variables peuvent être ou non représentées en trois dimensions, et elles disposent de différentes niveaux d'organisation : sélectif (permet de différencier les éléments) ; associatif (permet de confondre en un seul groupe les différents éléments) ; ordonné (permet de donner un ordre aux éléments) ; et quantitatif (ou métrique : permet de disposer d'une unité comptable). A travers ces niveaux d'organisation, les variables visuelles sont hiérarchisées en fonction de leurs qualités perceptives, résumant ainsi le pouvoir de chaque variable visuelle selon son efficacité (tab. 1-2).

Tableau 1-2 : Classement des variables visuelles. Source Bertin, 2005

		Association	Sélection	Ordre	Quantité
⇓	Dimension du plan	✓	✓	✓	✓
	Taille	X (Dissociation)	✓	✓	✓
	Valeur	X (Dissociation)	✓	✓	
	Grain	✓	✓	✓	
	Couleur	✓	✓		
	Orientation	✓	✓		
	Forme	✓			

En ce qui concerne la photographie aérienne (un système de signes polysémiques, comme déjà évoqué), son interprétation est rendue possible grâce à la présence simultanée de plusieurs variables (critères ou paramètres) qui représentent une zone ou un objet. Les variables prises en compte lors du processus de photo-interprétation sont : la forme, la taille, la teinte, la texture, l'arrangement (structure ou « pattern »), l'ombre et l'effet stéréoscopique (Chevalier, 1971 ; Gagnon, 1974 ; Weng, 2010). La forme permet de reconnaître les détails significatifs ; la taille permet d'évaluer les dimensions des

phénomènes, et de reconnaître les détails ; la teinte permet d'analyser l'état ou la nature de l'objet. La texture ou micro-arrangement des grains ; le « pattern » permettant de mettre en évidence l'organisation des espaces ; l'ombre donne de la profondeur à la photo sans faire appel au stéréoscope. Cependant, l'image, par elle-même, ne correspond qu'à une mesure relative de la luminance, exprimée par un ton de gris et son positionnement dans la dimension spatiale de la réalité. Les autres variables sont externes à l'image et elles intègrent les connaissances et les expériences apportées par l'interprète ou par l'expert (Caloz et Collet, 2001, Weng, 2010). Aussi, différentes clés d'interprétation sont-elles employées lors du processus de photo-interprétation selon : l'échelle[14], l'application, le niveau de détail, la finalité de l'étude menée, notamment. Nous trouvons donc : des études détaillées (sur photographies à grande échelle de 1 : 3 000 à 1 : 10 000) ; des études semi-détaillées (sur photographies à échelle de 1 : 10 000 à 1 : 20 000) ; et des études généralisées (à petite échelle, inférieure à 1 : 20 000) (Gagnon, 1974 ; Kraus et *al*., 1974).

En ce qui concerne **les études détaillées** à grandes échelles (de 1 : 3 000 à 1 : 10 000), Green (1956) a photo-interprété les logements sur des photographies aériennes à échelle 1 : 8 000. Il classifie les logements de 17 zones résidentielles à Birmingham (Etats-Unis) en différentes catégories : résidentielle (unifamiliale, multifamiliale, autres), commerciale, et industrielle (légère et lourde). Il s'est servi de plusieurs clés de photo-interprétation : la forme et la structure du toit, y compris les pignons, les lucarnes, les porches ; le nombre et la position des chenaux, ou des cheminées ; la taille, la forme et la hauteur du bâtiment ; la position du bâtiment par rapport à la rue ou à d'autres structures ; la présence de garages, abris d'autos, allées, véhicules et aires de stationnement ; la présence de trottoirs, sentiers, et l'entrée de promenades ; pour finir, la taille et la forme des lignes de démarcation des cours. Postérieurement, Hsu (1971) a photo-interprété des photographies aériennes à échelle 1 : 5 000 pour dénombrer les logements à Bolton et à Sandy Springs (Etats-Unis). Il utilise des cellules de la même dimension que celles des cartes pour marquer avec un point la présence des logements, ces cellules étant classées en : urbaine, rurale et semi-urbaine. Les clés d'interprétation

[14] L'échelle est le rapport de la distance mesurée sur le document (photographie ou carte) à la distance réelle mesurée sur le terrain (Chevalier, 1971).

sont, entre autres : la végétation à feuillage persistant, la couleur des toits et la présence d'allées.

S'agissant **des études semi-détaillées** (d'échelle 1 : 10 000 à 1 : 20 000), Lindgren (1971), après les travaux de Green (en particulier), estime les unités de logements à Boston (Etats-Unis) sur des photographies aériennes infrarouges à échelle 1 : 20 000. Il utilise comme clés dans la photo-interprétation : le type de toit, la taille de la structure, la présence de parkings, la division des constructions, le nombre d'étages, la quantité et la qualité de la végétation. Pour distinguer l'occupation du sol résidentiel/non résidentiel, les critères utilisés sont : la forme, la présence de parkings, la localisation relative, la quantité et la qualité de la végétation

Touchant **les études généralisées** (à échelles inferieures à 1 : 20 000), Tobler (1969) utilise une photographie (Apollo VI -AS6-2-1462) de la ville de Dallas, pour identifier et mesurer la zone bâtie de la ville, par le biais du rayon du cercle circonscrivant la tache urbaine. Avec ce même objectif de mesurer le rayon du cercle circonscrivant la tache urbaine, Anderson et Anderson (1973) ont conçu le système IDECS (*Image Discrimination Enhancement and Combination System,* en anglais) qui permet d'interpréter de manière automatique les zones urbaines/non urbaines d'une agglomération ; et ce afin d'estimer la population avec le modèle proposé par Nordbeck (§ 1.1.2). Dans sa recherche, 23 agglomérations du Kansas (Etats Unis) ont été photo-interprétées d'un côté, par cinq interprètes et, de l'autre, par le système mis au point. Les résultats obtenus montrent que les interprétations faites par le système IDECS étaient très proches de celles des interprètes. Postérieurement, Kraus *et al.* (1974) ont utilisé des photographies panchromatiques et infrarouges blanc et noir à petites échelles (1 : 60 000, 1 : 120 000 et 1 600 000). Ils ont photo-interprété l'utilisation du sol de quatre villes de Californie, en classant de la façon suivante : résidentiel unifamilial, résidentiel multifamilial, terrain de caravaning, et zone commerciale ou industrielle. Ultérieurement, Ogrosky (1975) utilise une photographie infrarouge à échelle 1 : 135 000 de la région du Puget Sound (Etats Unis). Il photo-interprète 18 agglomérations (classées préalablement en fonction de leur population et de leur taille) en mesurant : la zone bâtie, la longueur du réseau routier, et la distance entre les agglomérations. En 1977, Lo et Welch utilisent

les canaux rouge (5) et infrarouge (7), des images Landsat MSS prises entre 1972 et 1974 au-dessus de 13 villes de Chine, pour mesurer leurs zones bâties. Pour ce faire, ils créent des compositions colorées à échelle 1 : 500 000 et ils délimitent les taches urbaines par photo-interprétation.

L'interprétation visuelle est assez subjective car malgré les clés de photo-interprétation elle dépend du photo-interprète.

1.2.2 Le traitement d'image numérique

Le traitement d'image numérique est la *« mise en œuvre d'algorithmes de traitement destinés à extraire des informations significatives d'une image numérique prétraitée (CILF, 1997) ».* Ces algorithmes réalisent des analyses quantitatives sur les comptes numériques de l'image. Avec l'apparition de l'imagerie numérique, différentes méthodes de traitement de données ont été mises au point visant une exploitation optimale de l'information présente dans les images. Afin de comprendre les images satellite et les différentes méthodes de traitement de données, nous commencerons par les caractéristiques des images satellites ; nous poursuivrons avec les différents types d'amélioration et d'extraction des caractéristiques ; et nous finirons par la classification des images. Au cours de cet exposé nous ferons appel à différentes études réalisées dans le champ de la recherche géographique, notamment, dans la construction des variables pertinentes pour l'estimation de la population.

En ce qui concerne **les caractéristiques des images satellite**, celles-ci ont quatre résolutions de base : « spatiale », « spectrale », « radiométrique » et « temporelle » (Richards et Jia, 2006 ; Weng, 2010). La résolution spatiale détermine le niveau de détail avec lequel la surface est observée : cette information est représentée par la taille du pixel. Selon Puissant (2003), les images peuvent se classifier par leur résolution spatiale en : basse résolution (1000m), moyenne résolution (80m), haute résolution (10m-30m) et très haute résolution (<5 m). Cependant, étant donné les récentes améliorations dans la résolution spatiale (Orbview3 en 2003 avec 1m de RS; WordView 1 en 2007 avec 50cm RS ; GeoEye-1 en 2008 avec 41cm RS ; WordView 2 en 2009 avec 41cm RS ; entre autres),

nous pouvons définir la classification de la résolution spatiale de la manière suivante : basse résolution (de l'ordre du kilomètre - 1000m), moyenne résolution (de l'ordre de l'hectomètre - 100m), haute résolution (de l'ordre du décamètre - 10m) et très haute résolution (de l'ordre du mètre - 1m). La résolution spectrale correspond à la quantité de canaux -et à leurs longueurs d'onde- employée pour l'acquisition de l'image. On peut donc les subdiviser en : « mono-canal » (un seul canal), « multi-spectrale » (plusieurs canaux) et « hyper-spectrale » (plus de cent canaux). La résolution radiométrique fait référence à la capacité du satellite à discriminer la luminance, celle-ci se traduisant par le nombre de tons de gris possibles dans l'image. Nous trouvons assez fréquemment des images adont les pixels sont codés sur 8 bits (255 niveaux de gris), mais de plus en plus avec 16 bits (65535 niveaux de gris). La résolution temporelle indique le temps que le satellite met pour prendre une nouvelle image au même endroit.

S'agissant des **améliorations**, celles-ci visent à traiter les valeurs de façon à mieux exploiter l'information contenue dans celle-ci, c'est-à-dire à rendre l'information contenue dans l'image plus lisible pour l'interprète. Ainsi, ces améliorations peuvent être regroupées en trois groupes : **radiométrique, spatiale et spectrale** (Richards et Jia, 2006 ; Weng, 2010). D'un côté, les améliorations radiométriques visent à générer de nouvelles valeurs radiométriques du pixel par le biais du rehaussement du contraste (modification de l'histogramme) de l'image : ce traitement ne prend pas en compte le voisinage du pixel. De l'autre, les améliorations spatiales prennent en charge les modifications du détail géométrique de l'image, ces modifications impliquant une analyse du voisinage. Dans ce groupe d'améliorations, nous trouvons les différentes méthodes de filtrage (convolution, lissage, rehaussement, entre autres), l'analyse de Fourier, la détection de contours, et enfin la détection de propriétés géométriques comme la texture. Quant aux améliorations spectacles, celles-ci génèrent de nouveaux ensembles de composants de l'image ou « néo-canaux » à partir d'opérations arithmétiques ou statistiques entre les canaux originaux de l'image. Ces transformations visent plusieurs objectifs : construire les indicateurs synthétiques d'un phénomène ; réduire le nombre de données ; et/ou créer des variables avec une signification thématique. Font partie de ce groupe d'améliorations : l'analyse en composantes principales (ACP) ; les transformations

orthogonales (par exemple « *Tasselled Cap* ») ; l'arithmétique des canaux (addition, soustraction, multiplication et division) ; les indices spectraux (construits par division ou soustraction de canaux) dont font partie les indices de végétation.

Pour évoquer les travaux de recherche pour l'estimation de la population et l'application d'améliorations **spatiales** dans la construction des variables pertinentes dans cette estimation, on peut citer ceux de Liu *et al.* (2006). Cette recherche s'est fondée sur le calcul de la **texture** d'une image Ikonos, laquelle a été utilisée comme variable dans la modélisation statistique de la population par régression linéaire. La texture a été calculée sur les taches urbaines homogènes (notée HUP en anglais) du littoral de Santa Barbara (Californie) via trois approches. Ces HUP ont été photo-interprétées suivant un système de classification modifié d'Anderson III, créant ainsi cinq classes : faible densité de logement unifamilial, densité moyenne de logement unifamilial, haute densité de logement unifamilial, logements collectifs, et zones d'utilisation mixte commerciale-résidentielle. La première approche, « GLMC » *(grey level co-occurrence matrix,* en anglais), mesure pour chaque HUP les indices de texture suivants : l'énergie, l'entropie, le contraste, la corrélation, la variance, et l'homogénéité. Chaque indice a été mesuré en faisant varier la distance d'analyse entre 1 et 9 pixels. La deuxième approche, « semi-variance », construit un semi-variogramme isotrope expérimental pour chaque HUP en faisant varier la distance d'analyse entre 1 et 20 pixels, puis en créant un vecteur contenant ces semi-variances pour décrire la texture. La dernière approche, « mesures spatiales », calcule plusieurs descripteurs de texture fondés sur la distribution de trois groupes (surface bâtie, végétation et autres, classés au préalable) à l'intérieur de chaque HUP. Ainsi ont été calculés : le pourcentage de paysage (PLAND) de surface bâtie ; le PLAND de végétation ; la densité de la tache (PD) de surface bâtie ; la PD de végétation ; la distance euclidienne (ENN) entre les taches bâties et l'écart type de la ENN ; la cohésion ; l'agrégation ; et l'indice de diversité. Les données utilisées pour calculer les textures ont été le NDVI et le canal du proche infrarouge.

En ce qui concerne les améliorations **spectrales**, Harvey (2002) a modélisé la population de certains districts en Australie par le biais de la régression ; et ce en se fondant sur les variables construites grâces aux différentes transformations appliquées à une image Landsat TM (prise en 1988). Afin de mieux exprimer les différentes variables

utilisées par cet auteur, nous retiendrons comme notation b_i pour la moyenne des comptes numériques des pixels du canal i (i,j = 1,2,3,4,5,7) dans un district donné. De cette manière sont proposées les variables « arithmétiques » suivantes : a) $b_i * b_i$, b) $b_i * b_j$, c) b_i / b_j, et d) $(b_i - b_j)/(b_i + b_j)$. Dans ce travail, sont mesurées des variables comme : la variance, l'écart-type, et le coefficient de variation$(c = \frac{\text{écart-type}}{\text{moyenne}})$.

Plus récemment, en 2005, Li et Weng se sont servis des améliorations **spatiales et spectrales**, appliquées sur une image Landsat ETM+ de la ville d'Indianapolis (Etats-Unis), afin d'explorer leur corrélation avec les chiffres de la population. Pour ce faire, six groupes de variables de télédétection ont été utilisées. Le premier groupe a utilisé la réponse radiométrique de chaque canal dans chaque unité de recensement. Le deuxième groupe réalise l'analyse en **composantes principales** sur les six canaux de l'image, retenant seulement les trois premiers composants. Le troisième groupe se centre sur les **indices de végétation**, calculant ainsi : l'indice de la différence de végétation normalisé (NDVI), l'indice de végétation ajusté du sol (SAVI), l'indice de la différence de végétation ré-normalisé(RDVI), le NDVI transformé (TNDVI), l'indice de végétation simple (SVI), et une proportion simple (RVI). Le quatrième groupe analyse la mixture spectrale à l'intérieur de chaque unité de recensement. Le cinquième groupe aborde la texture par le moyen de la variance, laquelle a été mesurée sur les canaux 3 et 7 (respectivement rouge et proche infra-rouge) de l'image LandSat, faisant varier la grille d'analyse entre 3x3, 5x5 et 7x7. Le dernier groupe a utilisé le canal 6 pour mesurer la température dans chaque unité de recensement.

A propos des **indices de végétation**, Lo (1997) s'est servi de ceux-ci pour déterminer la qualité de vie dans un milieu urbain. Il a utilisé des images satellite Landsat d'Athens-Clarke (Géorgie – Etats-Unis) pour extraire l'occupation du sol, le NDVI, et la température ; les variables socio-économiques ayant été tirées du recensement. Ces informations ont été couplées à l'aide d'une ACP et de la superposition spatiale, et elles permettent de conclure que le NDVI est fortement lié à l'environnement morphologique et socioculturel. Lo avance ainsi l'utilité des images satellite comme complément du recensement afin de donner une perspective environnementale dans l'analyse urbaine.

Touchant la **classification**, celle-ci est définie par le Conseil International de la langue française (CILF, 1997) comme *« l'action, ou son résultat, consistant à répartir par classes, ou catégories, l'ensemble des éléments d'une scène »*. Cette répartition (regroupement ou zonage) permet soit de créer des zones où les pixels sont regroupés selon leur ressemblance spectrale, soit de déterminer les contours d'un groupe de pixels formant une région interprétable du point de vue thématique. Ce processus de répartition diffère du processus de photo-interprétation car dans ce dernier, le photo-interprète se sert de différents critères en simultané (§ 1.2.1) pour délimiter les zones. Dans le cas de la répartition par traitement numérique, celle-ci se fonde sur des critères de ressemblance entre pixels ou groupes de pixels (c'est-à-dire des régions).

Elle peut s'opérer selon trois stratégies : l'approche « par pixel », l'approche « par zone », et l'approche « par objet » (De Jong et Van Der Meer, 2005 ; Puissant, 2004). L'approche « par pixel » prend le pixel comme référence, se fondant ainsi sur la réponse spectrale. L'approche « par zone » utilise non seulement la réponse spectrale, mais intègre aussi des données auxiliaires, telles les données produites lors des améliorations spatiales (par exemple la texture et la détection de contours). L'approche « par objet » (ou orientée-objets) s'applique à des régions segmentées au préalable dans une image. En effet, dans ce type d'approche, la segmentation consiste en un prétraitement : différentes méthodes de segmentation de l'image existent, mais elles peuvent être rassemblées à leur tour en trois nouvelles approches (Blaschke *et al.*, 2005) : « pixel », « frontière » et « région ». La classification englobe de manière générale trois étapes (Schowended, 2007) : l'extraction des caractéristiques, l'apprentissage des pixels et l'étiquetage des pixels.

La première étape, l'extraction des caractéristiques, s'occupe de la préparation des données sur lesquelles la classification sera faite : par exemple les canaux originaux ou encore les canaux résultant de différentes améliorations.

La deuxième étape, l'apprentissage : c'est le processus par lequel les pixels sont sélectionnés pour entraîner le classificateur à reconnaitre les thèmes ou les classes souhaités. Lors de cette étape sont également déterminées des frontières (règles) de décision qui vont partitionner les caractéristiques en accord avec les propriétés des pixels apprises. Cette étape est donc soit dirigée par l'analyste, soit non dirigée ; et dans ce dernier cas, elle est réalisée à l'aide d'un algorithme informatique. Il existe différents

algorithmes tant dans l'apprentissage dirigé que dans l'apprentissage non dirigé. Parmi les algorithmes non dirigés, on peut citer (Girard et Girard, 1989 ; Richards et Jia, 2006) « l'identification des pics d'histogrammes », « la classification ascendante hiérarchique (CAH) », « la classification séquentielle par nuées dynamiques » où l'agrégation se fait autour de centres mobiles (k-moyennes, ISODATA). Quant aux algorithmes dirigés, ils sont générés par deux grands types de méthodes : « paramétrique », avec une hypothèse de distribution ; et « non paramétrique », sans hypothèse sur la distribution. Du côté des méthodes paramétriques des algorithmes dirigés, on peut citer (Girard et Girard, 1989 ; Richards et Jia, 2006) « la méthode de vraisemblance maximale», où le pixel à classer est affecté à la classe qui offre la probabilité la plus élevée. Cette méthode se fonde sur des règles de décision bayésiennes ou sur certaines modifications de celles-ci par exemple : « la règle de décision de vraisemblance maximale», « des modèles multi-variés de classe normale », « les surfaces de décision », ou « les seuils ». On peut citer également (Richards et Jia, 2006) la classification par « distance minimale » (celle-ci requérant un échantillon d'apprentissage moins important), et la classification par « la distance de Mahanalobis ». D'un autre côté, parmi les méthodes non paramétriques d'algorithmes dirigés, on trouve en particulier (Girard et Girard, 1989 ; Richards et Jia, 2006) : « la méthode de l'hyperboîte » (parallélépipédique), « l'arbre de décision », « les réseaux neuronaux », « le séparateur à vaste marge » (*Support Vector Machine,* en anglais), et « la discrimination linéaire ».

Lors de la troisième étape, l'étiquetage, les règles de décision pour partitionner les caractéristiques sont appliquées à l'image tout entière afin de classifier tous les pixels. Pour l'apprentissage dirigé, l'étiquette correspondant à la classe est déjà fixée ; tandis que pour l'apprentissage non dirigé, l'analyste doit attribuer des étiquettes aux partitions faites. L'étiquetage des pixels peut se faire en suivant des nomenclatures de classification déjà définies, par exemple : le système de classification d'Anderson (Anderson, 1971), le projet CORINE Land Cover (Commission de Communautés Européennes, 1993) ; ou simplement en utilisant un schéma de nomenclature proposé par l'analyste selon ses besoins.

Abordons les travaux de recherche pour l'estimation de la population et la classification d'images satellite, dans le but de construire les variables pertinentes dans

l'estimation de la population. Nous décrirons d'abord les méthodes de classification non dirigée, ensuite celles de classification dirigée, et enfin celles de classification orientée-objets.

Concernant la **classification non dirigée**, Iisaka et Hegedus (1982) ont classifié, en fonction de la densité de population, deux images Landsat MSS (1 et 2) avec l'algorithme ISODATA. Ultérieurement, en 2001, Lo a classifié la radiance des images DSMP avec la méthode de seuils naturels. Ces images mesurent la lumière nocturne émanant de la surface de la Terre : dans cette étude (afin d'estimer la population en Chine), les taches éclairées ont été groupées en six classes. Plus récemment, en 2009, Viel et Tran ont classifié une image LandSat ETM (sur les 6 canaux) avec l'algorithme ISODATA, afin d'obtenir les surfaces bâties/non bâties. Par ailleurs, ils ont aussi utilisé la classification dirigée avec la méthode de vraisemblance maximale, cette fois avec 7 classes, en fonction de la densité.

S'agissant de la **classification dirigée**, la méthode de vraisemblance maximale est la plus répandue (Richards et Jia, 2006). Tel est le cas pour l'étude de Lo (1992) qui a classifié une image SPOT1 (3 canaux) en 12 classes, bien qu'il en ait seulement retenu deux *in fine* : la « résidentielle à haute densité », et la « résidentielle à basse densité »; estimant ainsi que le canal proche infrarouge est un excellent estimateur de la densité de population. Postérieurement, Harvey (2002b) utilise cette même méthode sur des images Landsat TM, choisissant également 12 classes dont une urbaine et les 11 autres réservées à d'autres usages. En 2006, Langford a classifié avec cette méthode les 6 canaux d'une image Landsat (prise le 12 juin 1992). La classification a porté dans un premier temps sur 25 classes ; mais par la suite, il en ressort une image classifiée en trois catégories : rurale, sous urbaine et urbaine dense. Plus récemment, en 2010, Dong a classifié deux images Landsat TM avec la même méthode de vraisemblance maximale. La classification a porté sur les 12 canaux des deux images, et les regroupent en cinq classes : résidentielle, commerciale, industrielle, zone de transport, et surfaces de végétation/eau. Une autre méthode utilisée est la classification dirigée par « les seuils ». Cette méthode s'est appliquée aux images DSMP afin de classifier l'intensité de la lumière nocturne (Dobson

et al., 2000 ; Brigss, 2006). Dans ces deux travaux, d'autres variables, notamment l'occupation du sol, ont été utilisées pour modéliser la population.

Touchant la classification **orientée-objets**, Lu et son équipe (2010) se sont servis d'une image Quickbird (4 canaux R, V, B, PIR avec 2,8m de résolution spatiale, et 1 canal panchromatique avec 0,7m résolution spatiale) et des données LiDAR (*Light Detection And Ranging,* en anglais) (densité de 2,3 pts m^2) afin d'identifier les bâtiments. Trois classifications sur ces données ont été opérées : la première n'a porté que sur l'image ; la deuxième n'a pris en compte que les données LiDAR ; et la dernière a inclus et l'image et les données LiDAR. Sur chacune de ces trois classifications chaque canal a été pondéré de la même manière, mais subjectivement. La segmentation des images a été réalisée avec l'approche fractale d'évolution nette (FNEA - Li *et al*., 2008) disponible sur le logiciel Defiens2007. Cette méthode est multi-niveau : elle met au point les objets, en zoomant de façon croissante, en partant du pixel. Cette méthode requiert plusieurs valeurs de paramètres spécifiées par l'utilisateur comme : la pondération des données d'entrée, la proportion couleur/forme, la proportion aspect lisse/compacité, et le facteur d'échelle. De plus, quatre niveaux d'analyse ont été testés pour le facteur d'échelle 10. La validation de ces niveaux s'est faite de manière visuelle sur l'image QB et les données LiDAR. En définitive, le niveau d'analyse 30 a été retenu car il offrait le meilleur niveau de segmentation, maintenant un compromis satisfaisant entre l'homogénéité des objets et le nombre des objets obtenus. Néanmoins, chaque bâtiment est composé d'un certain nombre de parties, variant de 2 à 4. Enfin 35 paramètres ont été choisis pour identifier les bâtiments en utilisant l'image Quickbird et les données LiDAR. Parmi ceux-ci, 25 correspondant à l'image Quickbird, 10 correspondant aux données LiDAR.

1.2.3 La modélisation des données de hauteur des objets

Une autre ligne de force dans les progrès de la télédétection réside dans les avancées de la modélisation du relief ; la photogrammétrie comme la télédétection ayant évolué, les outils comme les techniques pour l'extraction du relief ont suivi cette même tendance. Après l'apparition pour le grand public en 1851 (Bajac, 2010) de la « photographie stéréoscopique », l'évolution de la modélisation s'est poursuivie avec l'utilisation du

« radar », appareil qui émet des impulsions en hyperfréquences et reçoit l'écho de ces impulsions (mot universellement adopté depuis 1945 - CILF, 1997). Ce type d'instruments peut être aéroporté, ou embarqué à bord d'un satellite. SEASAT[15], lancé en 1978, est le premier satellite de télédétection civile à comporter un capteur Radar à synthèse d'ouverture (RSO). A présent, il existe différents satellites équipés d'instruments radar permettant d'obtenir des informations à échelles variées. Parallèlement, autour de 1961, les travaux sur le laser pour satellite ont commencé à Goddard[16] : le premier satellite de type exploratoire, Explorer-Beacon B, équipé d'un laser a été mis en orbite à Cap Kennedy en octobre 1964 (Vonbun *et al.*, 1977). Plus récemment, en janvier 2003, était lancé GLAS[17] (*Geoscience Laser Altimeter System,* en anglais) le premier instrument LiDAR pour l'observation globale continue de la Terre. Le « LiDAR » utilise comme source émettrice le laser (CILF, 1997) et il permet de modéliser le relief : l'utilisation du LiDAR aéroporté est la plus fréquente grâce à sa précision (Popescu *et al.*, 2011). Toutefois, l'utilisation du relief - notamment la hauteur des bâtiments- s'est intégrée relativement tard aux recherches géographiques visant l'estimation de la population et nous pouvons dire que la modélisation de la hauteur est récente dans la modélisation de population : on le constate par le faible nombre de recherches qui l'ont intégrée.

Ainsi, en 2008, Wu et son équipe ont calculé la hauteur des bâtiments via les données LiDAR (0,61m) en soustrayant la hauteur au sol de la hauteur au toit. Ils ont également vérifié par photo-identification l'existence des contours de bâtiments à l'aide de photographies aériennes. Enfin, ils ont calculé le volume de chacun des bâtiments pour modéliser la population.

Postérieurement, en 2010, Dong et son équipe ont également utilisé des données LiDAR afin de générer les modèles numériques du terrain et de la surface (respectivement MNT - MNS) avec la méthode d'interpolation inverse de la distance (IDW) sur une grille de 1m*1m. Ils ont couplé ces modèles avec la classification des images satellites : des analyses spatiales de la surface et du volume ont été calculées pour modéliser la population.

[15] http://ilrs.gsfc.nasa.gov/satellite_missions/list_of_satellites/seas_general.html (consulté le 16 janvier 2012)
[16] Centre spatiale de la NASA http://www.nasa.gov/centers/goddard/about/index.html, (consulté le 16 janvier 2012)
[17] http://glas.gsfc.nasa.gov/ (consulté l 16 janvier 2012)

De même pour Lu et ses collègues (2010) qui ont utilisé ce type de données pour l'extraction des bâtiments. Ils ont produit des MNT et MNS à partir d'un réseau de triangles irréguliers (TIN, *triangular irregular network,* en anglais). Deux approches ont été utilisées pour modéliser la population à travers des modèles de régression: l'approche « surface », fondée sur la surface des zones « bâtiments résidentiels » ; et l'approche « volume », toujours fondée sur la surface mais également sur la hauteur des bâtiments.

Plus récemment, Lwin et Muraya (2011) se sont servis, également, des données LiDAR pour calculer la hauteur moyenne des bâtiments (MNS – MNT, construits à travers des TIN), et ainsi modéliser la population par une approche volumétrique.

Tableau 1-3 : Utilisation des images pour construire des variables pertinentes dans l'estimation de la population : résumé

Utilisation des images pour construire des variables pertinentes dans l'estimation de la population : résumé			
Méthode	« interprétation visuelle »	« traitement d'images numériques »	« modélisation des données de hauteur »
Données d'entrée	Photographie aérienne Images satellite	Images satellite	Photographie aérienne Données radar Données laser
Description	Lecture et analyse des images, fondée sur les capacités psycho-visuelles d'une personne	Extraction des informations significatives d'une image numérique à partir d'algorithmes de traitement	Utilisation des données afin de modéliser le relief
Approches	Photo-interprétation	Classification d'images : par pixel, par région, orientée-objets	Stéréoscopie Interpolation spatiale

1.3 La télédétection au service de l'épidémiologie

L'utilisation de la télédétection en l'épidémiologie est plus récente, même si l'épidémiologie s'est servie depuis 1792 (Barret, 2000) des apports de la géographie, particulièrement de la cartographie. Dès la fin des années soixante, les épidémiologistes ont découvert l'intérêt de l'exploitation de données issues de la télédétection,

notamment en vue d'associer les indices d'apparition de la maladie avec les caractéristiques de l'Homme et de son environnement (Cline, 1970). En 1970, Cline propose un éventail d'utilisations possibles en épidémiologie pour les différents types d'images existant à l'époque, mais il envisage aussi des améliorations dans l'imagerie, et leurs applications à différentes maladies ; son travail a lancé différentes pistes de recherche. A l'heure actuelle, un grand nombre d'études épidémiologiques utilisent la télédétection, et nous regrouperons ces applications en trois groupes : l'étude des maladies infectieuses, la mesure des expositions, et l'estimation de la population.

En premier lieu, **l'étude des maladies infectieuses** : selon les conclusions des travaux de Hay (1997) et Herbreteau *et al.* (2007), elle représente la plupart des applications de télédétection en épidémiologie. L'application de la télédétection se fonde sur le croisement des mesures du rayonnement électromagnétique (REM) avec les mesures d'une maladie et de son vecteur[18] (Curran *et al*., 2000, Herbreteau *et al.*, 2007 ; Kalluri *et al.*, 2007). Selon Curran *et al*. (2000), ce croisement se fait autour de trois éléments-clés : l'identification de la couverture du sol ; la distribution spatiale de l'habitat d'un vecteur ; et la distribution spatiale de la maladie à transmission vectorielle. D'ailleurs, certains auteurs (Hay, 1997 ; Bonne *et al*., 2000 ; Curran *et al*., 2000) considèrent la couverture du sol comme une variable déterminante dans les études épidémiologiques, caractérisée par télédétection. D'autres auteurs (Beck *et al.*, 2000 ; Manguin et Boussinesq, 1999) incluent en plus de ces critères : le suivi, la surveillance, la cartographie des risques et la prédiction des maladies à transmission vectorielle (King *et al.*, 2004 ; De La Rocque *et al.*, 2004). Globalement, les traitements ou les analyses issus de la télédétection les plus demandés sont la couverture et l'occupation du sol, l'indice de végétation et le calcul de la température (Herbreteau *et al.,* 2007). Ce fait est établi par des recherches plus récentes (Ford *et al.*, 2009 ; Tran *et al.*, 2010 ;Viel *et al.*, 2011).

[18] « Arthropode hématophage assurant la transmission biologique active d'un agent pathogène, d'un vertébré à un autre vertébré » (Rhodain et Perez, 1985)

S'agissant de la **mesure des expositions**[19], la télédétection permet aux épidémiologistes d'explorer les relations entre l'environnement et la santé. Cela concerne au premier chef l'épidémiologie environnementale, centrée autour de la problématique suivante : pourquoi et comment les facteurs environnementaux affectent-ils la santé humaine ? (Merril, 2008). Ainsi, parmi ces explorations, on trouve les travaux sur la mesure de l'exposition aux pesticides (Maxwell *et al.*, 2010 ; Ward *et al.*, 2000), qui examinent la relation entre les pesticides utilisés sur les cultures à proximité des résidences individuelles et les retombées sanitaires. Ces deux recherches ont utilisé des images Landsat : l'une pour calculer une série «d'indices de végétation (NDVI) » multi-temporelle comme un indicateur de l'état des cultures, ce qui peut améliorer les modèles existants de mesure de cette exposition en Californie (Maxwell *et al.*, 2010) ; l'autre pour générer une carte historique d'utilisation du sol dans l'étiologie du lymphome non-hodgkinien au Nebraska (Ward *et al.*, 2000). D'autres auteurs focalisent ces recherches sur la mesure de l'exposition à la pollution atmosphérique**,** notamment les particules fines (Hu, 2009 ; Hu et Rao, 2009 ; Paciorek et Liu, 2009) ; et ce dans le but d'examiner l'association des maladies coronariennes avec ces polluants. Ces recherches ont utilisé des images MODIS (avec une résolution spatiale de 1,1km) pour modéliser la concentration des polluants. Une autre branche de la recherche en épidémiologie environnementale s'intéresse à la mesure de l'exposition aux vagues de chaleur, notamment à une température élevée la nuit (Laaidi *et al.,* 2012). Cette étude met en évidence le fait que l'exposition à une température élevée la nuit sur plusieurs jours augmente la probabilité de décès au cours d'une vague de chaleur en milieu urbain. Des images AVHRR-NOAA (avec une résolution spatiale de 1,1km) ont été utilisées pour calculer les températures à Paris et dans le Val-de-Marne durant l'été 2003.

A notre connaissance, peu de recherches épidémiologiques ont impliqué la télédétection pour **estimer** les taux d'incidence de maladies (et donc **les populations** exposées). On peut se référer à la recherche menée en 2002 par Tran et son équipe, dans le but de cartographier l'incidence de la fièvre Q dans les environs de Cayenne, en Guyane

[19] « Une exposition peut représenter une exposition réelle (par exemple, un produit chimique toxique ou un microorganisme), un comportement (par exemple, le lieu où l'on travaille ou socialise), ou un attribut de la personne (par exemple, l'âge, le sexe, la race) » (Merril, 2008)

française, en utilisant les données de télédétection. Ces auteurs sont partis des images SPOT pour cartographier la couverture du sol, et ensuite générer un indice de densité de la population, lequel a servi pour déterminer les zones de haute incidence de la maladie pendant la période 1996-2000. Une comparaison avec les effectifs du recensement a été faite pour valider les résultats obtenus, tant pour la densité de population que pour le taux d'incidence de la maladie, montrant une forte corrélation entre les deux. De plus, ils ont constaté que toutes les zones définies comme « à haute incidence » selon le recensement ont été également identifiées par les données télédétection. Cela leur a permis de démontrer le potentiel indéniable de la télédétection comme outil de calcul rapide des taux d'incidence d'une maladie, dans les enquêtes épidémiologiques. La principale limite de cette approche est d'aboutir à des taux d'incidence relatifs et non absolus. Plus récemment, en 2009, Viel et Tran (§ 1.1.3) ont développé un modèle de population fondé sur des images Landsat ETM+, reproductible et exportable, conduisant à l'estimation de l'incidence absolue (et non relative) pour de petites zones géographiques, et palliant ainsi l'absence d'effectifs de population pour l'analyse épidémiologique. Leurs recherches élargissent ainsi la gamme des utilisations possibles envisagées en 1970 par Cline et réaffirmées par Hay en 1997, qui associe les avancées futures en recherche épidémiologique aux progrès éventuels apportés par la télédétection dans la modélisation de la population.

Tableau 1-4 : Résumé de l'utilisation des images dans la recherche épidémiologique

Utilisation des images dans la recherche épidémiologique : résumé			
Type de recherche	Etude des maladies à transmission vectorielle	Mesure d'une exposition, afin d'explorer les relations avec la santé	Estimation de la population, dénominateur des taux
Apport de la télédétection	Distribution spatiale de l'habitat d'un vecteur ; et celle de la maladie.	Caractérisation de l'environnement	Modélisation de la population
Résolution spatiale des images	Moyenne	Moyenne – Basse	Moyenne – Haute

L'épidémiologie s'est servie de la télédétection à résolutions différentes et pour différentes recherches.

« Il faut donc, avant toute chose, définir, en fonction de la problématique posée, les éléments qu'il est envisageable d'extraire de l'image, puis les méthodes qui vont permettre cette extraction » C. Weber (Images satellitaires et milieu urbain, 1995)

Chapitre 2. Besançon, ville d'art et d'histoire, mais aussi ville verte : un cadre diversifié propice pour une méthodologie reproductible

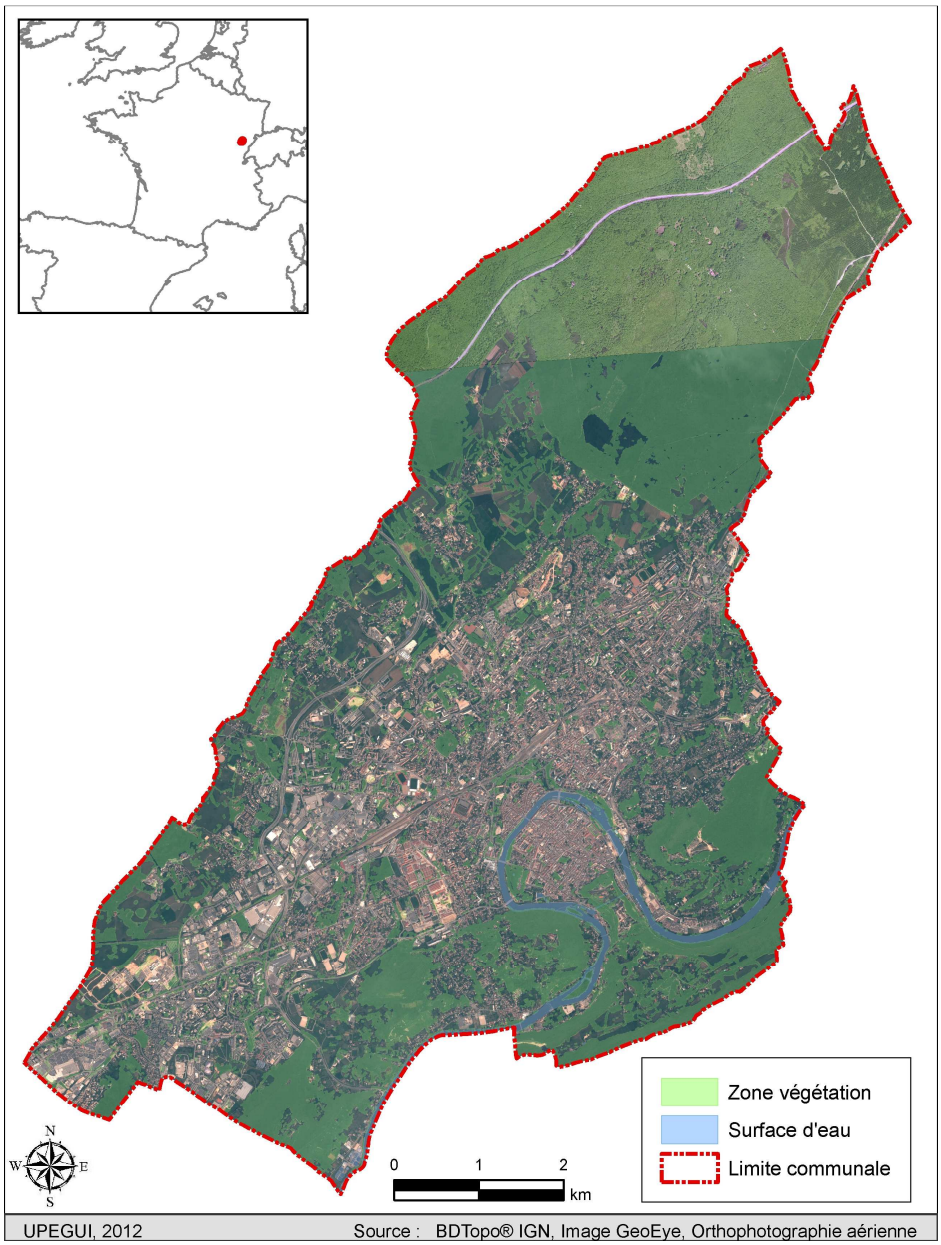
La ville de Besançon (fig. 2-1), est devenue française en 1674, puis capitale de la Franche-Comté en 1677 (Bidalot, 2009). La richesse de son architecture, comme de son patrimoine historique et culturel, lui permet de bénéficier du label « Ville d'art et d'histoire[20] » depuis 1986, et de faire partie du « Patrimoine mondial de l'humanité » depuis 2008 (La Citadelle, fort Griffon et enceinte fortifiée). Besançon est aussi, d'après l'étude réalisée par l'agence BMJ CoreRatings en 2004, la première ville verte de France avec une densité d'espaces verts entretenus de 204m²/hab. (Reverchon, 2004).

Besançon se situe sur la bordure occidentale du faisceau bisontin, c'est-à dire sur le bourrelet pentu que limite la montagne jurassienne, dans le prolongement du Revermont et du faisceau salinois (Courtieu, 1982). La ville est traversée par le cours du Doubs qui forme un méandre dans laquelle la ville ancienne s'est bâtie. Elle s'étend sur 65,05km², à une altitude variant de 235m à 620m (Bidalot, 2009). Sept collines sont bien identifiées : Bregille, La Citadelle (nommée auparavant Saint-Etienne), Chaudanne, Rosemont, Planoise, Roche d'Or et Fort-Benoit.

---

[20] « Le ministère de la Culture et de la Communication assure depuis 1985, dans le cadre d'un partenariat avec les collectivités territoriales, la mise en œuvre d'une politique de valorisation du patrimoine et de sensibilisation à l'architecture, concrétisée par l'attribution du label « Ville ou Pays d'art et d'histoire ». » - http://www.vpah.culture.fr/label/label.htm (consulté le 24 février 2012).

Figure 2-1 : Besançon, localisation générale

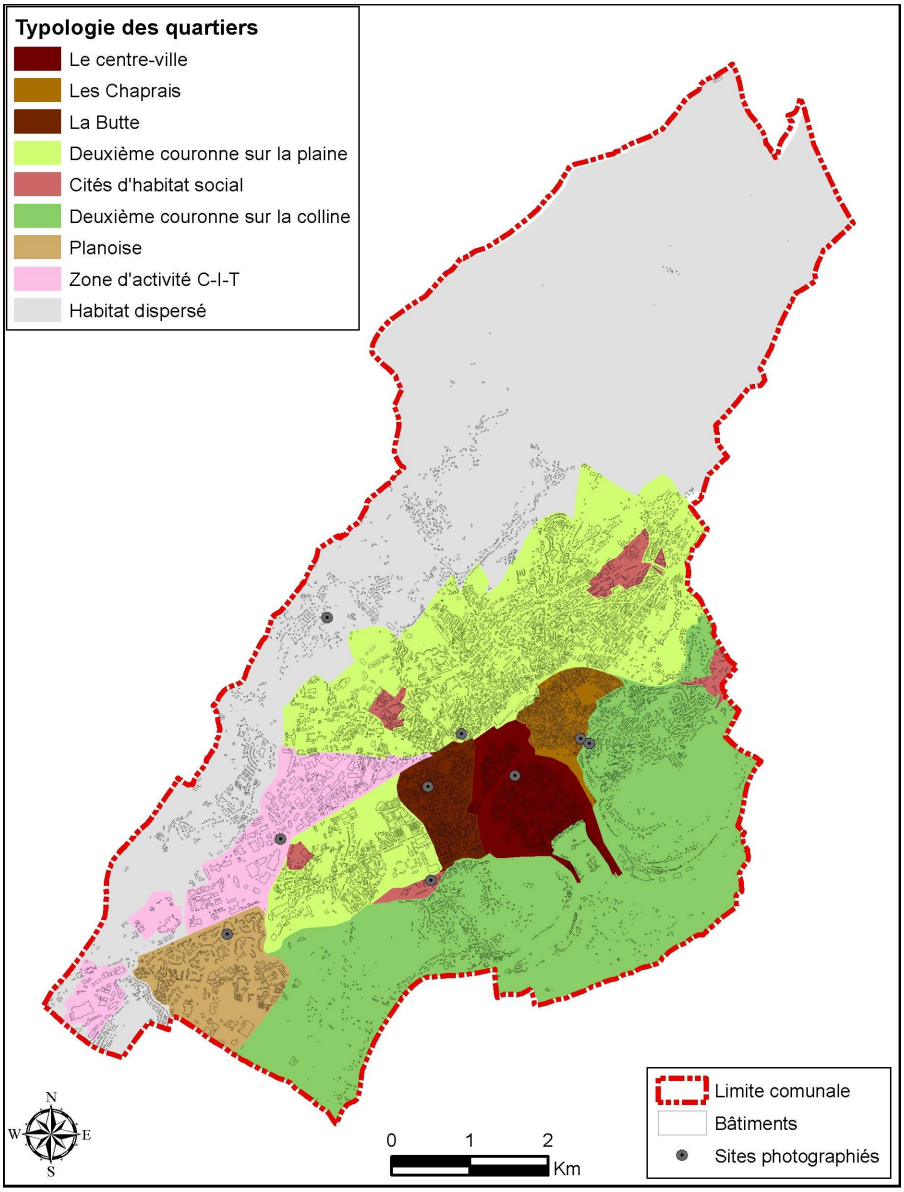


2.1 Des quartiers aux bâtiments

L'expansion spatiale des bâtiments à Besançon est liée à l'évolution historique de la ville, conduisant ainsi à différentes enveloppes ou couronnes. Afin de décrire la typologie générale des bâtiments bisontins, nous nous appuierons sur la typologie de quartiers proposée par Houot (1999, fig. 2-2) qui sera enrichie à partir du Plan Local d'Urbanisme de Besançon (D.U.H., 2007). La variété des typologies de bâtiments dans les quartiers, ainsi que la mixité de leur localisation et de leur distribution rendent difficile

leur différenciation et l'extraction « bâti/non bâti ». Cette difficulté dans l'extraction s'accroît encore avec les différentes combinaisons possibles des variables visuelles clés : la forme, la taille, la teinte, la texture, l'arrangement (structure ou « pattern »), l'ombre et l'effet stéréoscopique, dans l'identification des objets (§ 1.2.1), notamment des bâtiments.

Figure 2-2 : Typologie des quartiers, adaptée de Houot (1999)



2.1.1 Le centre-ville

Bien que la fondation de Besançon se perde dans le temps (Bidalot, 2009), on peut considérer que son histoire commence en 58 avant notre ère avec l'occupation de la ville par César (Courtieu, 1982). La ville ancienne s'est construite tout au long des siècles, notamment du XIII^{ème} au XVIII^{ème} siècles. Cependant, après la conquête de Besançon par Louis XIV, Vauban exerça une forte influence et transforma le centre ville en modernisant les défenses, en introduisant les vastes quadrilatères des casernes et surtout en redessinant entièrement les bordures et les accès de la ville (D.U.H., 2007). A l'heure actuelle, le centre-ville de Besançon, selon le D.U.H., possède les caractéristiques de la ville européenne « classique ». Il est donc compact et resserré avec un bâti continu, ponctué de bâtiments publics monumentaux, organisé le long de rues et de places, et délimité par les anciens remparts (fig. 2-3).

Figure 2-3 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) du centre-ville

2.1.2 Première couronne : entre 1850 et 1950

Une première couronne, correspondant au développement de la ville entre 1850 et 1950, est caractérisée par un urbanisme de densité moyenne avec un bâti semi-continu, fait essentiellement de maisons individuelles, de petites cités-jardins ou petits immeubles organisés surtout à partir de voies et chemins préexistants (D.U.H., 2007).

2.1.2.1 Les Chaprais

Dans ce quartier, la compacité, la densité, et l'organisation peuvent être vues comme proches de celles de la ville ancienne (D.U.H., 2007) bien qu'elles en différent quelque peu. Différents types de construction conservant les cœurs d'ilots, placés sur un réseau de voies assez dense sont présents dans ce quartier : il est constitué par un grand nombre de maisons individuelles, de villas, et de résidences.

Figure 2-4 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) des Chaprais

2.1.2.2 La Butte

Ce quartier se caractérise par la présence de maisons individuelles avec des îlots où la végétation domine. Même s'il existe aussi des immeubles collectifs, l'ensemble reste peu dense.

Figure 2-5 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) de La Butte

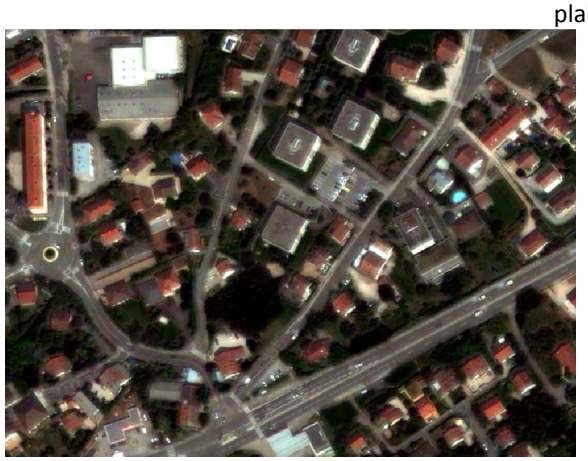
2.1.3 Deuxième couronne : Après Guerre

Après la seconde Guerre Mondiale, la forte croissance démographique (taux de natalité élevé, exode rural, immigration) et la pénurie de logements conduit à un net changement d'échelle dans l'urbanisation (D.U.H., 2007). Ainsi, au début des années 1950, les premières cités de grande ampleur prennent corps, si bien que le tissu urbain s'étend et se densifie avec des immeubles collectifs qui ferment les espaces occupés par des maisons individuelles. La mixité entre bâtiments collectifs et maisons individuelles définit le nouveau paysage de ces secteurs de la ville. Bien entendu, les immeubles construits augmentent en taille par rapport à ceux des siècles précédents.

2.1.3.1 Sur la plaine

Une partie de cette deuxième couronne s'est construite sur la plaine de la ville, c'est-à-dire dans la partie la plus extérieure du méandre du Doubs mais qui reste toutefois parallèle au cours de la rivière. Cette frange de la couronne est relativement dense tout en restant verte.

Figure 2-6 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) de la deuxième couronne sur la plaine



2.1.3.2 Cités d'habitat social

Dans cette expansion de la ville sur la plaine, les cités d'habitat social trouvent leur place sur d'anciens terrains militaires, ou sur de grands terrains. Ces cités correspondent à

de grands immeubles (en surface et en hauteur) avec des toits-terrasse. Nous trouvons ainsi : la cité de Palente, de Montrapon, de Brulard, des Clairs Soleil et les tours de l'Amitié.

Figure 2-7 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) des cités d'habitat social

2.1.3.3 Sur la colline

Une autre partie de la couronne se construit vers la partie (intérieure) sud du méandre et sur la colline. Une mixité s'observe également sur cette partie de la couronne, mais avec d'autres caractéristiques dues au relief. Ici prédomine l'habitat dispersé, avec des maisons individuelles qui se mêlent à quelques immeubles collectifs de faible hauteur.

Figure 2-8 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) de la deuxième couronne sur la colline

2.1.4 Dernière enveloppe : ville récente

Bien que surgie Après Guerre, cette dernière enveloppe présente des caractéristiques différentes de celles de la deuxième. Précisons que la construction de maisons individuelles ne s'est pas ralentie durant l'expansion de la ville. A la différence du travail de Houot (1999), nous scinderons en deux groupes la typologie de quartier « zones industrielles et habitat périphérique dispersé » car la typologie des bâtiments y est clairement différenciable.

2.1.4.1 Planoise

Ce nouveau quartier s'est construit dès la fin des années 1960 et pendant les décennies 1970 et 1980 dans le but d'accueillir 40 000 habitants (Houot, 1999). Dans un premier temps, ces bâtiments collectifs correspondent à des immeubles de grande hauteur, relativement denses, et très uniformes. Ensuite, un réaménagement des espaces a été opéré avec la poursuite de la construction dans ce nouveau quartier : on y place des immeubles d'habitation de hauteur moyenne, plus variés sur le plan architectural (Houot, 1999).

Figure 2-9 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) de Planoise

2.1.4.2 Zone d'activité commerciale, industrielle et tertiaire (C-I-T)

Le tissu industriel et commercial est largement concentré en périphérie, où des espaces étaient disponibles. L'activité commerciale s'est répartie vers les bordures, à l'ouest de la ville, d'où l'émergence de Châteaufarine. De même, la zone industrielle s'est

prolongée de Trépillot vers le bas des Tilleroyes. Pour finir, le nouvel hôpital J. Minjoz a entraîné avec lui des cliniques privées, constituant un pôle santé dans ses alentours (D.U.H., 2007). Ces bâtiments présentent des variations d'échelles (grandeur/hauteur) et de formes ; de là un paysage particulier pour cette partie de la ville.

Figure 2-10 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) de la zone d'activité C-I-T

2.1.4.3 Le périurbain : un habitat dispersé

Des zones pavillonnaires ont trouvé place autour de Planoise mais aussi dans la zone périurbaine du reste de la ville, constituant ainsi une « frange » urbaine avec densité faible, au bâti éparpillé ou à peine regroupé le long des voies (D.U.H., 2007) : le paysage naturel prédomine dans cette frange.

Figure 2-11 : Sous-image GeoEye (à gauche) et typologie du bâti (à droite) de l'habitat dispersé

2.2 Typologie des toitures

Nous nous intéressons aux toitures parce que, en télédétection, c'est à travers elles que nous approchons les bâtiments. Au delà des « bâtiments classés » avec des toitures remarquables témoignant d'une richesse architecturale (avec des époques et des styles différents), nous pouvons classifier les toitures en deux groupes selon qu'elles sont en pente ou en terrasse. Cette typologie nous semble importante car il a été déjà démontré par Lhomme (2005) qu'il existe une dépendance de la réponse radiométrique aux angles de visée, d'éclairement et aux inclinaisons des cibles. Nous trouverons donc des réponses radiométriques différentes sur une même toiture en fonction de l'inclinaison et de l'orientation de ses pans ; de ses accidents : lucarnes, lanterneaux, cheminées et autres ; ainsi que de ses matériaux.

2.2.1 Les types de toit

2.2.1.1 Les toitures en pente

Les toitures en pente se caractérisent par leurs pans inclinés : les formes les plus simples peuvent avoir entre un et quatre versants (Calvat, 2003) (fig. 2-12). De plus, une variante assez fréquente des formes simples correspond à un petit versant de forme généralement triangulaire -nommé « la croupe »- situé à l'extrémité des longs pans. Une autre variation fréquente est « le comblé » constitué par la couverture et la charpente avec des pans d'inclinaison différente.

Figure 2-12 : Formes simples de toits extraits de l'image GeoEye. De gauche à droite, toit à un seul versant ; à deux versants ; à quatre versants ; croupe droite ; et le comblé

2.2.1.2 Les toitures en terrasse

Les toitures en terrasse (fig. 2-13) -ou plates- se caractérisent par leurs pans horizontaux (Calvat, 2003) qui peuvent être, généralement, de tôles profilées ou de dalles de béton. Ce type de couverture est peu répandu pour les maisons individuelles mais fréquent en habitat collectif. La bordure d'une toiture-terrasse est délimitée par un acrotère, muret en béton, avec une certaine hauteur.

Figure 2-13 : Exemples de deux toitures en terrasse extraits de l'image GeoEye

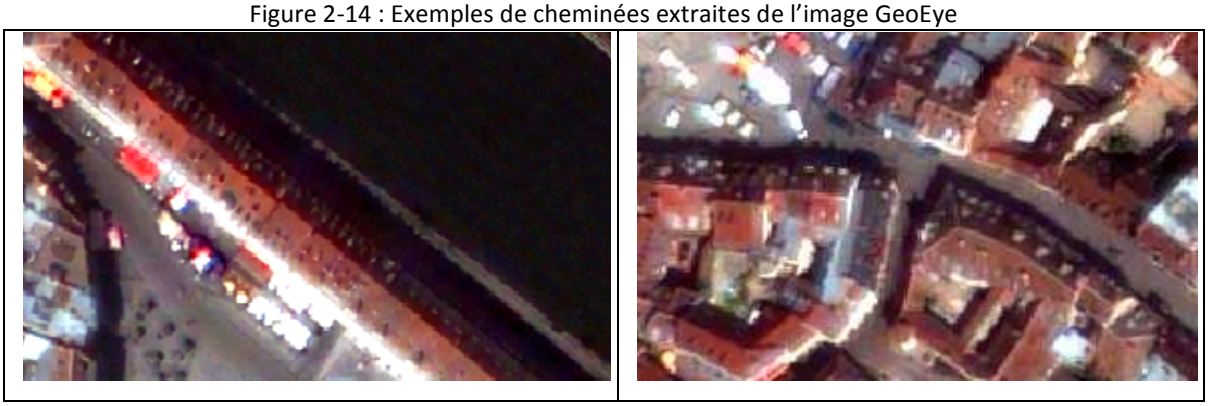
2.2.2 Les accidents de toitures

Selon Calvat (2003), l'accident de toiture désigne tout élément qui dépasse de la toiture tels que les souches de cheminée, les sorties de cheminée, ou autres. Nous incluons également dans cette catégorie les ouvertures du toit telles que les lucarnes, la chatière, le châssis à tabatière, le vasistas, la fenêtre de toit, la verrière, les lanternes d'aération, entre autres.

2.2.2.1 La cheminée : souche et sortie

La cheminée constitue un élément important dans les bâtiments, du moins dans les régions dites « tempérées ». Sur les images satellite à THRS, nous percevons la partie du conduit à fumée dépassant du toit du bâtiment qui apparaît comme un élément de petite taille visible sur la toiture (fig. 2-14). Ce conduit est appelé « souche de cheminée » s'il est maçonné ; sinon, s'il est métallique par exemple, il sera appelé « sortie de toit » (Calvat, 2003).

Figure 2-14 : Exemples de cheminées extraites de l'image GeoEye



2.2.2.2 Les ouvertures de toit

Les ouvertures de toit sont des « baies » ou des ouvrages édifiés sur la toiture afin de l'éclairer, de l'aérer, et/ou de donner accès au comble (fig. 2-15). Il en existe différents types selon leur forme mais aussi selon leur position, mais nous ne les détaillerons pas. Pour notre objectif (l'extraction de bâtiments), nous nous contenterons de savoir qu'elles jouent un rôle important dans la physionomie de la toiture, qu'elle soit en pente ou en terrasse, ou encore qu'elle recouvre une maison individuelle ou un habitat collectif. Ces ouvertures se traduisent par des variations dans la texture, le « pattern », le couleur, et la forme. En outre, elles peuvent générer des ombres sur la toiture en fonction de leur orientation.

Figure 2-15 : Exemples d'ouvertures de toit extraits de l'image GeoEye

2.2.3 Les matériaux de couverture

Afin de décrire de manière globale les différents types de matériaux de couverture prédominants à Besançon, nous nous appuierons sur l'étude menée en 1979 par l'ancien service du Délégué Régional à l'Architecture et l'Environnement (DRAE), qui propose quatre grands groupes de matériaux présents dans la région de Franche Comté. De plus, nous avons croisé ces sources documentaires avec celles désignées dans le travail de Chamouton (2009) sur « les maisons comtoises ». Bien que nous ne puissions pas distinguer la spécificité de ces matériaux, celle-ci se traduira sur une image à THRS par des différences de textures et de « patterns », ainsi que de couleurs (fig. 2-16).

Figure 2-16 : Exemples des matériaux différents extraits de l'image GeoEye. De gauche à droite ardoise, tuiles et béton

2.2.3.1 Les matériaux traditionnels

Dans ce groupe on trouve les matériaux présents dans la région comme : les petites tuiles plates ou écailles, les tuiles canal et l'ardoise.

2.2.3.2 Les matériaux originaux

Ce groupe se caractérise soit par sa création comtoise, soit par son utilisation particulière dans la région. Appartiennent donc à ce groupe les tuiles en terre cuite de différents type comme : l'emboîtable, la « villa », la double écaille, les tuiles violon, la plate-emboîtable, la double plate et la triple plate. Sans oublier les bardeaux de bois (le tavaillon), et les laves ou les dalles de grès.

2.2.3.3 Les matériaux substituts

Cette catégorie correspond aux matériaux modernes qui remplacent les matériaux traditionnels tout en gardant leur style. Nous trouvons dans ce groupe : « l'épiplaque », l'ardoise « ETERNIT » brune, la tuile-béton « prestige », les tuiles emboîtables et plates « super-Beaucourt » et « Valois », les couvertures en métal (tôles, aluminium, inox), et les plaques d'amiante-ciment.

2.2.3.4 Les matériaux banalisants ou peu durables

Dans cette catégorie se trouvent les matériaux banals comme les tuiles « mécanique standard », et les matériaux peu durables comme les tuiles « béton peintes ».

2.3 Données de référence

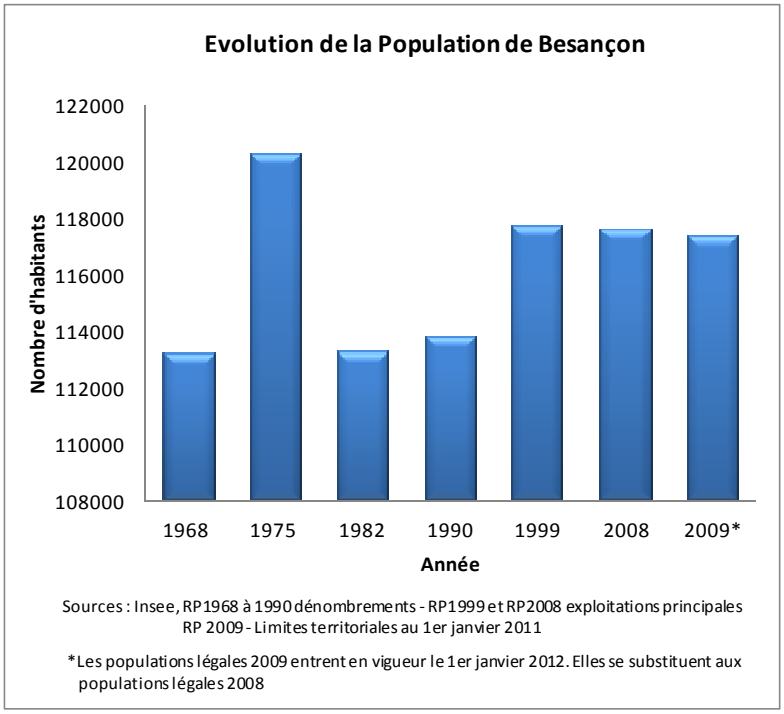
Pour la réalisation de notre recherche nous nous appuyons sur des données dites de « référence ». Ces données nous permettront de mettre au point et de valider la méthodologie proposée. Nous aborderons leur description, avec la même approche que celle utilisée dans « L'état de la question », c'est-à-dire en suivant les trois lignes d'intérêt de cette recherche : modélisation de la population, extraction des bâtiments, et calcul des taux d'incidence.

2.3.1 Effectifs de la population à Besançon - IRIS1999

Besançon, selon le recensement de 2009, compte 117 392 habitants (INSEE, 2009) en légère décroissance par rapport aux 117 691 dénombrés en 1999 (INSEE, 1999) (fig. 2-17). Cela représente une variation relativement faible (-0.25%), qui marque une stabilité de la population dans cette dernière décennie. D'ailleurs la dernière croissance de la population, à Besançon, a été enregistrée dans la dernière décennie du siècle dernier, et correspond chronologiquement à la création de la ville récente (Planoise). Or, le recensement de la population a fait l'objet d'une rénovation en 2004 (Article 20 du décret n° 2003-485 du 5 juin 2003) qui induit différentes difficultés, et qui privilégie, pour

l'instant[21], les évolutions de la population par rapport au recensement de 1999 (INSEE, 2010). Donc, dans cette recherche nous avons choisi de travailler avec les données du recensement de 1999.

Figure 2-17 : Evolution de la population de Besançon de 1968 à 2009



L'objectif principal de cette recherche étant d'estimer les taux d'incidence au niveau infra-communale (pour une aide à la décision en santé publique), nous avons pris comme données source les « ilots[22] » et les « Ilots Regroupés pour l'Information Statistique » (IRIS[23]), spécifiquement ceux de 1999 (ILOTS1999 et IRIS1999 respectivement).

[21] « Les populations légales sont désormais actualisées chaque année. Toutefois, les enquêtes de recensement étant réparties sur cinq années, il est recommandé de calculer les évolutions sur des périodes d'au moins cinq ans. Pour l'instant, **la référence pour le calcul des évolutions reste donc le recensement de 1999.** » http://www.insee.fr/fr/ppp/bases-de-donnees/recensement/populations-legales/commune.asp?annee=20 09&depcom=25056. [Consulté le 2mars 2012]

[22] L'îlot était l'unité géographique de base pour la statistique et la diffusion des recensements de la population jusqu'à celui de 1999. Les îlots étaient définis par l'Insee en concertation avec les communes.
- En zone bâtie dense : l'îlot représentait le plus souvent un pâté de maison, éventuellement scindé en cas de limite communale ou cantonale traversant le pâté de maison ;
- En zone « périphérique » : l'îlot était un ensemble limité par des voies (ou autres limites visibles) découpant cette zone en plusieurs morceaux.
Les îlots pouvaient être vides d'habitants (par exemple une gare).
http://www.insee.fr/fr/methodes/default.asp?page=definitions/ilot.htm [consulté le 14 avril 2012]

[23] Les IRIS correspondent au découpage du territoire en mailles de taille homogène (visé de 2 000 habitants), réalisé par l'INSEE afin de préparer la diffusion du recensement de la population de 1999. Depuis leur création, les caractéristiques démographiques de certains IRIS ont pu évoluer, il est donc utile de spécifier leur année de référence.

2.3.2 Bâtiments -BDTopo® 2006

Les bâtiments représentent les cibles que nous cherchons à extraire dans la première étape de notre procédure (fig. 2-1). Ils occupent 7,73% de la surface totale de la ville (surface calculée par ArcGIS - système d'information géographique). Afin de valider notre méthode, nous nous sommes servis de la base de données BD TOPO®, notamment de la thématique « Bâti », comme « vérité terrain ». La BD TOPO® fait partie du « Référentiel géographique à Grande Echelle » (RGE®) de précision métrique constitué par l'Institut Géographique National (IGN, 2009).

S'agissant du « Bâti », la BD TOPO® possède différentes précisions géométriques planimétriques variant de 0,5m à 10m, selon les données d'origine. De même, il existe une variation de la précision géométrique altimétrique : avec des valeurs inférieures à 1 m ou inferieures à 2,5m ; ou sans aucune valeur altimétrique. En outre, la thématique « Bâti » est constituée de différentes classes : bâti indifférencié, bâti remarquable, bâti industriel, construction légère, cimetière, piste d'aérodrome, réservoir, équipement sportif de plein air, construction linéaire, construction ponctuelle, et construction surfacique. A leur tour, ces classes peuvent se subdiviser en différents types de bâtiments. Bien que cette base de données soit « officielle », elle présente certaines limites quant à la modélisation géométrique. Les bâtiments de moins de 20m² en sont exclus. Les bâtiments d'une surface comprise entre 20 et 50m² n'y figurent que s'ils sont situés à plus de 100m d'une habitation. La BD TOPO® intègre le bâti du cadastre, même si la géométrie des bâtiments de cette base de données n'est pas exactement superposable avec la géométrie des bâtiments du cadastre. Seules les cours intérieures de plus de 10m de large sont représentées par un trou dans la surface bâtie. Le contour extérieur du bâtiment est le plus souvent celui du toit, et dépend donc de la prise de vue de la photographie sur laquelle il a été restitué. Dans le cas de plusieurs bâtiments contigus, ceux-ci sont généralement considérés comme un seul et même objet, et ainsi seul le contour extérieur est saisi. Quant à la modélisation de la hauteur des bâtiments, les bâtiments contigus ne seront considérés comme deux objets différents, que s'ils manifestent une différence de hauteur supérieure à 10m ou une surface supérieure à 400m². Au delà de ces contraintes, nous avons retenu comme « bâtiments de référence », ceux appartenant aux classes « bâti indifférencié », « bâti remarquable », « bâti

industriel » avec également un caractère permanent, cela excluant ainsi les tours, donjons, moulins, tribunes, serres, et silos.

2.3.3 Mesures d'incidence des cancers issues du Registre général des tumeurs du Doubs

La seule source de données, en France, pour mesurer l'incidence des cancers est constituée des registres des cancers. Le Registre des tumeurs du Doubs (RTD) est l'un des plus anciens registres généraux des cancers en France. Il a été créé en 1977, et étendu en 2007 au Territoire de Belfort (Woronoff et Danzon, 2008). Le RTD a comme principales sources de données[24] (Woronoff et Danzon, 2008) :

- les laboratoires d'anatomie et de cytologie pathologiques qui transmettent, ou mettent à disposition du registre, les comptes rendus des prélèvements effectués chez les patients ;

- les établissements de santé publics et privés qui transmettent, ou mettent à disposition du registre, des informations provenant des dossiers médicaux, soit par le biais des données du Programme Médicalisé des Systèmes d'Information (PMSI), transmises par les Départements de l'Information Médicale (DIM), soit directement par les services hospitaliers ou les réunions de concertation pluridisciplinaires ;

- les caisses des différents régimes d'assurance-maladie qui transmettent les listes de patients bénéficiant d'une Affection Longue Durée n°30 (ALD 30).

Ainsi, le RTD assure un recueil continu et exhaustif de données nominatives dans la population résidant dans le département au moment du diagnostic -quelque soit le lieu de prise en charge- à des fins de santé publique et de recherche.

En dehors de données publiées sur le site Internet du RTD, l'accès aux données recueillies est restreint et soumis à la Charte d'utilisation des données du Registre des tumeurs du Doubs, dans le cadre de protocoles de recherche[25]. Cet accès respecte les clauses légales énoncées par la Commission nationale de l'informatique et des libertés (CNIL) demandant le cryptage des données individuelles sur l'accès aux données médicales nominatives. Enfin, il est possible d'avoir accès aux données agrégées

[24] http://www.chu-besancon.fr/registretumeursdoubs/ (consulté le 12 mars 2012)
[25] http://www.chu-besancon.fr/registretumeursdoubs/CharteUtilisateurDonnees0706.pdf (consulté le 12 mars 2012)

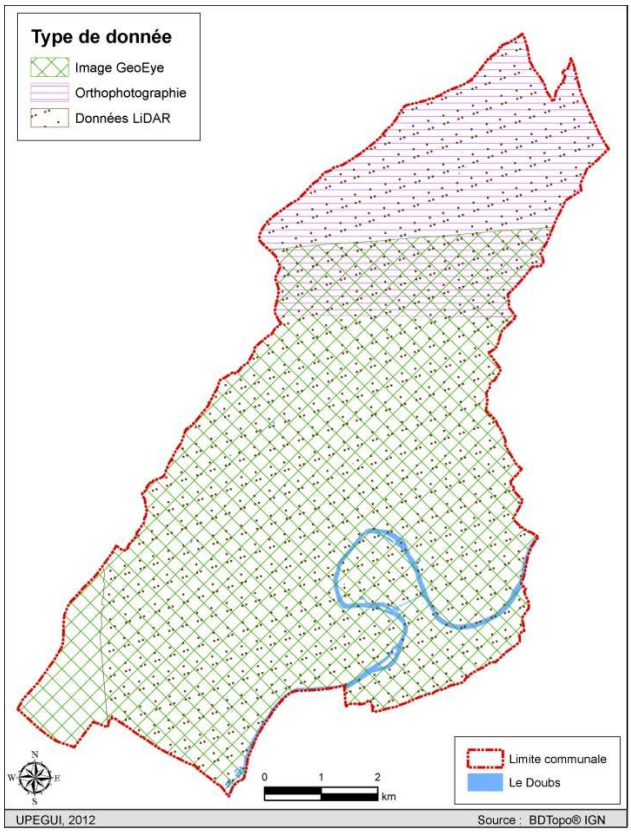
anonymisées, lesquelles sont géo-référencées indirectement (code de la commune, adresse des patients) permettant, après un travail de reformatage, un géocodage au niveau de l'IRIS[26]

In fine, les données épidémiologiques dont on dispose pour mener cette recherche correspondent à 222 cas de lymphomes non hodgkiniens diagnostiqués à Besançon entre 1980 et 1995 (Floret *et al.*, 2003), et à 491 cas de cancer du sein (chez les femmes) diagnostiqués à Besançon entre 1996 et 2002 (Viel *et al.*, 2008).

2.4 Données issue de la télédétection

Dans ce sous-chapitre nous exposerons les données télédétection dont nous disposons pour extraire les bâtiments (un pilier fondamental de cette recherche), selon différentes méthodes (cf. chapitre 3). Toutefois, ni leurs caractéristiques, ni leur couverture sur la zone d'étude, ne sont égales (fig. 2-18), ce qui pose certaines contraintes, exposées dans le chapitre « Extraction des bâtiments ».

Figure 2-18 : Couverture des données, issues de la télédétection, sur Besançon



[26] http://www.sante-environnement-travail.fr/minisite.php3?id_rubrique=1027&id_article=4539, (consulté le 12 mars 2012)

2.4.1 L'image GeoEye

GeoEye-1[27] est un satellite commercial à très haute résolution spatiale, mis sur orbite (héliosynchrone) le 6 septembre 2008 à une altitude de 681km. GeoEye-1 fonctionne par le biais d'un appareil-photo électro-optique qui prend des images panchromatiques avec 0,50m[28] de résolution spatiale ; et des images multispectrales avec 2m de résolution spatiale et 4 canaux. Les longueurs d'onde sont précisées dans le tableau ci-dessous :

Tableau 2-1 : Résolution spectrale du satellite GeoEye-1

Canal	**Longueur d'onde**
Panchromatique	450 - 800 nm
Multispectrale	
Bleu	450 - 510 nm
Vert	510 - 580 nm
Rouge	655 - 690 nm
Proche infrarouge	780 - 920 nm

L'image GeoEye dont nous disposons pour mener cette recherche a été acquise le 7 août 2009. Elle est la propriété de l'UMR CNRS 6249 « Chrono-Environnement ». Nous disposons de l'image panchromatique et de l'image multispectrale avec ses 4 canaux (système de référence, Projection UTM Nord fuseau 32 - Système géodésique WGS84). Les conditions de la prise de vue sont détaillées dans le tableau suivant :

Tableau 2-2 : Caractéristiques techniques de la prise de vue de l'image acquise sur Besançon

Caractéristique	**Degrés**
Azimut nominal de capture	137,06
Élévation nominale de capture	74,89
Angle d'azimut du soleil	149,68
Angle d'élévation du soleil	55,99

[27] http://www.geoeye.com/CorpSite/assets/docs/brochures/GeoEye-1_Fact_Sheet.pdf

[28] 50 cm est la résolution spatiale avec laquelle les images sont livrées en raison des politiques de sécurité des Etats-Unis, mais celles-ci sont au départ acquises avec 0,41cm de résolution spatiale. http://www.geoeye.com/GeoEye101/satellite-imagery/collection-method.aspx (consulté le 13 mars 2012).

L'image panchromatique couvrant la zone d'étude compte 25 145*22 061 pixels et présente une taille de 1,19Go, tandis que l'image multispectrale compte 6 287*5 516 pixels et occupe 348,14Mo. L'image fusionnée (à savoir 4 canaux avec 0,5m de résolution spatiale) correspond à 25 145*22 061 pixels et 4,61Go.

2.4.2 L'orthophotographie aérienne

Une orthophotographie est une « *image photographique sur laquelle ont été corrigées les déformations dues au relief du terrain, à l'inclinaison de l'axe de prise de vue et à la distorsion de l'objectif » … « Toutefois, elle peut présenter des déformations résiduelles d'autant moins négligeables que les pentes de terrain sont fortes et les superstructures plus nombreuses et plus élevées (CILF, 1997)* ». Les photographies servant à cette orthophotographie ont été acquises lors d'une campagne aérienne effectuée le 07 mai 2009, avec une caméra numérique UltraCam-Xp (Vexcel/Microsoft Corp). Ces images ont été prises avec une résolution spatiale de 0,20m (14 001*10 001 pixels et 585,38 Mo de taille) et quatre canaux avec différentes longueurs d'onde (Scholz et Gruber, 2009) (tab. 2-3). Néanmoins, des adaptations par rapport au contraste des clichés ont été effectuées afin d'obtenir la mosaïque finale (Société Aerodata France, 2009). En outre, le système de projection de l'orthophotographie est le système Lambert II Etendu (méridien de Paris et non celui de Greenwich). Par ailleurs, cette orthophotographie est propriété de la MSHE Claude Nicolas Ledoux-USR 3124 CNRS.

Tableau 2-3 : Résolution spectrale de la camera UltraCam-Xp

Canal	**Longueur d'onde**
Bleu	400 - 570 nm
Vert	480 - 630 nm
Rouge	580 - 700 nm
Proche infrarouge	690 - 1000 nm

2.4.3 Données LiDAR

De même que l'orthophotographie, les données LiDAR sont la propriété de la MSHE Claude Nicolas Ledoux (Nuninger *et al.*, 2010), mais les missions d'acquisition de données LiDAR ont eu lieu les 8 - 10 avril 2009. L'acquisition des données a été réalisée

avec le scanner LMS-Q560 (RIEGL), avec une densité de 10,6 pts/m², et une précision altimétrique inférieure à 20cm (SAF, 2009). Toutes les trajectoires ont été calibrées et géoréferencées (position et altitude) à partir d'un système de positionnement global (GPS), et de mesures inertielles prises pendant les vols. Cela permet de déterminer avec une grande précision la position (latitude, longitude et altitude) et l'orientation (tangage, roulis et lacet) des centres de perspective de chaque cliché au moment de l'exposition.

Le modèle numérique des surfaces (MNS) et le modèle numérique du terrain (MNT) ont été produits à partir des données de LiDAR, avec une résolution spatiale de 50cm, en utilisant la méthode d'interpolation « distance pondérée ». Cette interpolation a été faite avec « BCAL LiDAR Tools[29] », plugiciel « Open source » pour ENVI développé par le laboratoire aérospatial de Boise Center (BCAL) de l'Université d'Idaho State. De plus, les dimensions de chacun des modèles est 20 730*27 132 pixels et leur taille atteint 2,27Go.

Nous avons exposé dans ce chapitre un aperçu de la diversité des typologies de bâtiments présentes dans la ville de Besançon, de même que de leurs toitures car les bâtiments seront approchés à travers elles. De plus, nous avons présenté les différents types de données dont nous disposons pour mener à terme cette recherche, cela afin de cerner les difficultés que nous affronterons lors de l'extraction des bâtiments, et afin de déterminer la portée de cette étude.

[29] http://code.google.com/p/bcal-lidar-tools/

« Il n'y a pas de solutions simples à des problèmes complexes. Il y a en général plusieurs approches. Aucune n'est complète, mais chacune est susceptible d'apporter un éclairage particulier »R. Caloz et C. Collet (Analyse spatiale de l'information géographique, 2011)

Chapitre 3. Extraction des bâtiments à partir de données télédétection : du pixel à la reconnaissance des formes

La première étape de la méthodologie proposée concerne l'extraction des bâtiments, afin de construire la variable « bâti » / « non bâti ». Cette variable permettra de désagréger l'information spatiale du recensement, et donc d'estimer la population. Elle fait partie, selon la classification des variables (physiques ou socio-économiques) pertinentes dans l'estimation de la population (§ 1.1.3), de la catégorie « occupation / utilisation du sol ». Néanmoins, nous avons visé à extraire les « bâtiments », au lieu des « surfaces », grâce à la THRS des données de télédétection que nous utilisons.

Pour l'extraction des bâtiments, nous utilisons quatre approches différentes parmi le grand éventail de choix de la classification, qui vont des approches dites « classiques » (fondées sur le pixel) jusqu'à des approches plus « récentes », orientées-objets, entrant ainsi dans le domaine de la reconnaissance de formes.

Afin de rendre les résultats obtenus comparables entre eux (visuellement) et donc plus compréhensibles, un échantillon (fig. 3-1) par typologie de quartier (§ 2.1) a été choisi (fig. 3-2).

Figure 3-1 : Localisation des échantillons à détailler

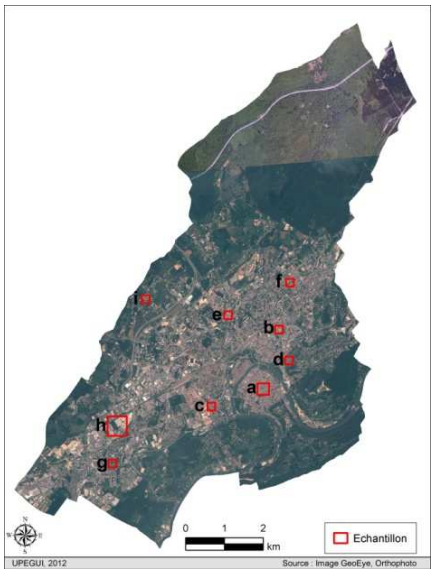
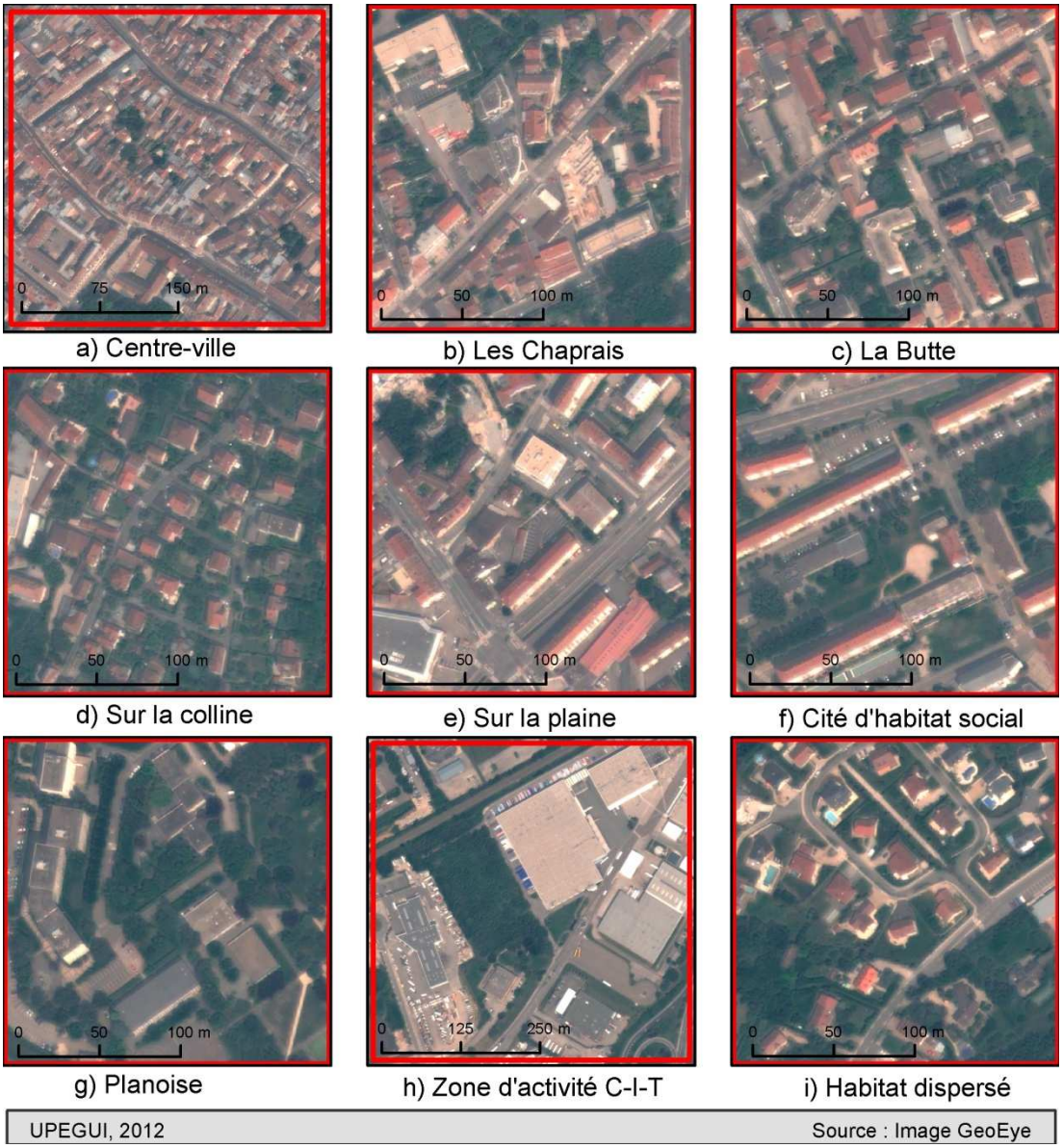


Figure 3-2 : Echantillon par typologie des quartiers



En outre, des prétraitements -corrections géométriques[30] et traitements radiométriques[31]- ont été effectués sur les différentes données, pour les intégrer dans le même système de référence (système de coordonnées projetées sur NTF - Lambert II Etendu) et avec la même taille du pixel, à savoir 50cm. Par ailleurs, sur toutes les images classifiées, un filtre « majorité » a été appliqué afin de les nettoyer la variabilité des pixels isolés.

3.1 Classifications fondées sur le pixel

En ce qui concerne cette stratégie de classification, tant le processus d'apprentissage « dirigé » que celui « non dirigé » ont été testés.

3.1.1 ISODATA : une méthode non dirigée

ISODATA (*Iterative Self-Organizing Data Analysis Technique*, en anglais - Ball et Hall, 1965) a été la première méthode que nous avons utilisée pour extraire le « bâti/non bâti » à partir des images THRS. Ce choix se justifie par deux raisons. La première concerne directement notre objectif qui vise à évaluer l'apport des données de THRS par rapport aux données de HRS (comme Landsat) (Viel et Tran, 2009) dans le même cadre urbain, Besançon. En effet, c'est dans la recherche de Viel et Tran (2009), que la méthode d'ISODATA a fourni les meilleurs résultats dans l'estimation de la population. La seconde raison est plus liée à une finalité de santé publique (prévision et action sanitaire), où l'estimation de la population doit être : simple, automatique, reproductible, et exportable. ISODATA est l'une des techniques les plus utilisées dans le traitement d'image numérique car elle est d'une applicabilité facile. Ce fait est constaté à travers le grand nombre de publications qui l'incluent (Manakos *et al.*, 2000 ; Melesse et Jordan, 2002 ; Sohn et Dowman, 2007 ; Zambon *et al.*, 2012 ; entre autres), ainsi qu'à travers la diversité

[30] « Correction appliquée à une image, ou traitement d'une donnée, en vue d'en réduire les erreurs de nature géométrique et de la rendre superposable à une carte » (CILF, 1997). Dans notre cas, l'image GeoEye a été orthorectifiée avec les rcp (*Rational Polynomial Coefficients*, en anglais) et le MNT. Pour le MNT une mosaïque a été faite avec le MNT issu des données LiDAR et celui de l'IGN (pour la partie sans couverture des données LiDAR).

[31] Notamment la dégradation de l'orthophoto, c'est-à-dire « la fusion, par calcul de moyennes radiométriques de pixels contigus disposés en carrés, effectuée dans l'intention d'élargir la limite de résolution au sol d'images numériques » (CILF, 1997).

des thématiques qui l'utilisent : agriculture, astronomie, géomorphologie, océanographie, parmi d'autres. Avec ISODATA, l'utilisateur n'a pas de connaissances *a priori*, les classes étant automatiquement créées par le logiciel. Néanmoins, la difficulté de cette méthode, et en général des méthodes non dirigées, réside dans l'identification thématique (l'étiquetage) *a posteriori* des classes spectrales issues du processus de classification, c'est-à-dire la production du sens à partir de la classification.

Cette méthode itérative, ISODATA, regroupe les valeurs radiométriques de l'image dans un nombre de classes défini par l'utilisateur. L'hypothèse sur laquelle s'appuie cette méthode est la suivante : la probabilité d'un pixel d'appartenir à une classe donnée augmente avec la diminution de l'écart entre le barycentre de la classe et ce pixel. Ainsi, l'agrégation de tous les pixels (à classifier) au barycentre le plus proche, se fait en fonction d'une distance spectrale euclidienne, suivant quatre grandes étapes (fig. 3-3) :

1- choix initial du nombre de classes et choix des barycentres initiaux de ces classes ;

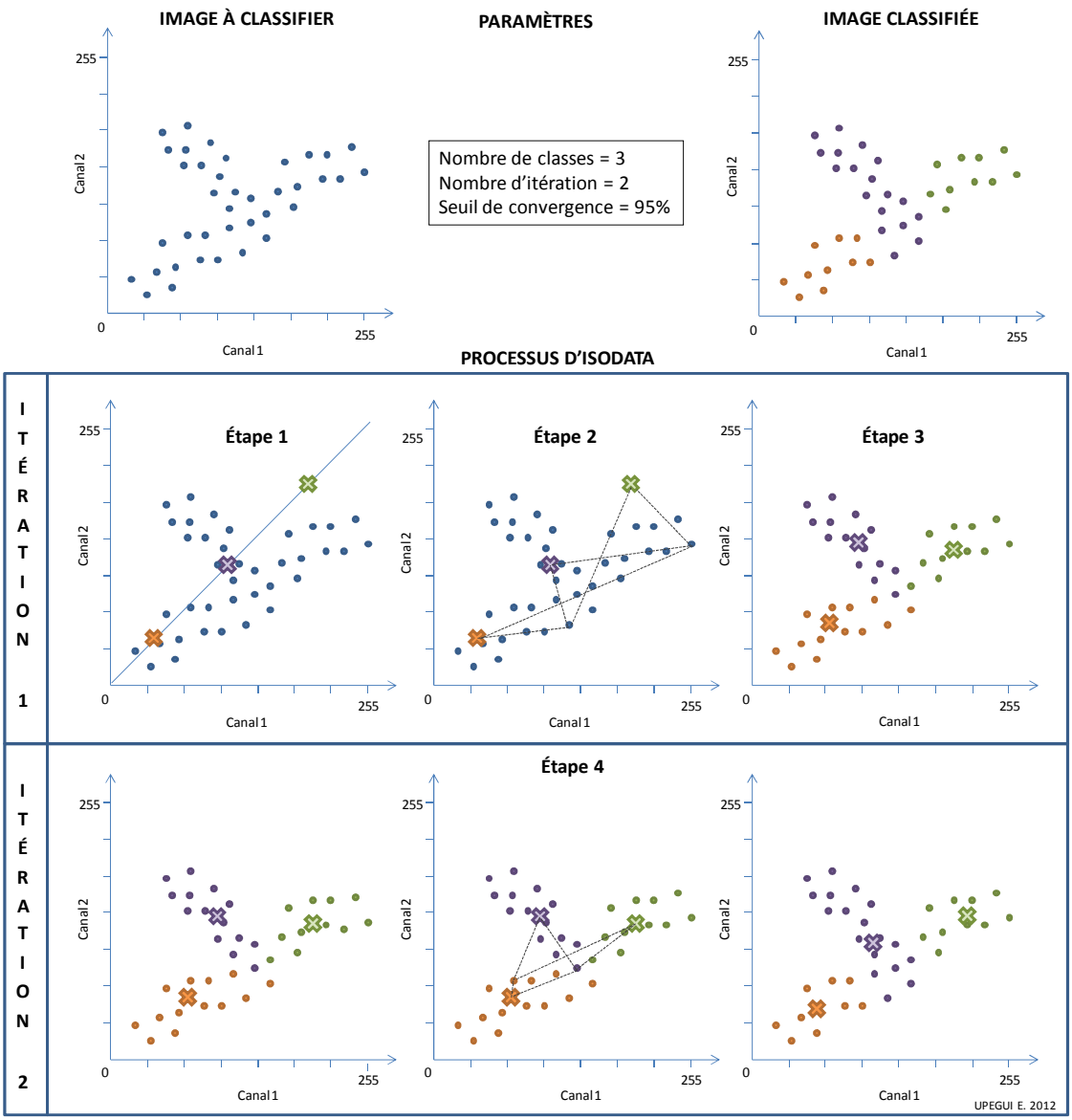
2- calcul de la distance entre chaque pixel et chacun des barycentres ;

3- affectation de chaque pixel au barycentre le plus proche, et re-calcul de la moyenne du barycentre en fonctions des pixels affectés ;

4- répétition des étapes 2 et 3 jusqu'à convergence.

Aussi, l'utilisateur doit-il spécifier au minimum le nombre de classes à établir, le nombre d'itérations à réaliser, et un seuil qui permet d'arrêter éventuellement le processus si le résultat devient stable (si le processus « converge », c'est-à-dire le moment où les pixels ne fluctuent pas entre classes entre une itération et l'itération suivante). De plus, ISODATA peut regrouper deux classes si leurs barycentres sont plus proches qu'un certain seuil, ou si le nombre de pixels dans une classe est plus petit qu'un nombre donné. En revanche, une classe peut être séparée si son écart-type dépasse une certaine valeur, ou si elle contient un nombre de pixels très supérieur au seuil minimal de pixels.

Figure 3-3 : Illustration du processus d'ISODATA



L'application de cette classification à notre zone d'étude est représentée dans la figure 3-4. Nous avons classifié de manière indépendante l'image GeoEye de l'orthophotographie aérienne, parce qu'elles ont des caractéristiques spectrales différentes (tab. 2-1 et tab. 2-3). L'image GeoEye et l'orthophotographie aérienne ont chacune 0,50m de résolution spatiale et 4 canaux multispectraux (fig. 3-5a). Pour obtenir une seule image classifiée « bâti/non bâti », une mosaïque a ensuite été faite. Etant donné la taille des images (4,61Go pour GeoEye, et 585,38Mo pour l'orthophoto), le temps de calcul pour les traiter, et la redondance d'information que ces images peuvent présenter du fait d'avoir trois canaux (sur quatre) captés dans la longueur d'onde du « visible », nous avons pris deux décisions. Premièrement, celle de procéder à l'extraction

des caractéristiques (première étape de la classification), qui consiste en la création des « néo-canaux » résultant d'une amélioration spectrale par l'ACP (à savoir ACP1, ACP2, ACP3, ACP4 - fig. 3-5b). Deuxièmement, celle d'améliorer la classification (fig. 3-5c) par la technique des masques[32] emboités, qui permet d'identifier les bâtiments, par approximations successives. En outre, pour l'étiquetage (troisième étape de la classification) des classes « bâti/non bâti », nous nous sommes appuyée sur l'interprétation de la « signature spectrale[33] » (fig. 3-5d) de chacune des classes, laquelle a été calculée à partir des canaux originaux de l'image (R-V-B-PIR). Ainsi, dans une classification X_i, l'analyse de la signature spectrale de chaque classe permet de discriminer les surfaces classifiées comme « non bâties ». Un masque (fig. 3-5e) excluant ces zones non bâties est alors créé et une nouvelle classification (X_{i+1}) est opérée sur la nouvelle zone d'analyse définie. C'est pour cela que cette méthode devient « dépendante de l'opérateur », bien qu'elle soit non dirigée.

La classification a d'abord porté sur les deux composantes les plus significatives de l'ACP (tab. 3-1). Puis les autres composantes de l'ACP ont été injectées progressivement : l'ACP3 a été introduite dans la 4ème classification, puis l'ACP4 a été glissée dans la dernière classification. Au total, cinq classifications et quatre masques ont été nécessaires pour obtenir l'image finale classifiée (fig. 3-5f). Dans chacune des classifications, les paramètres ont été les suivants : nombre de classes = 20 ; nombre d'itérations = 60 ; et seuil de convergence = 95%.

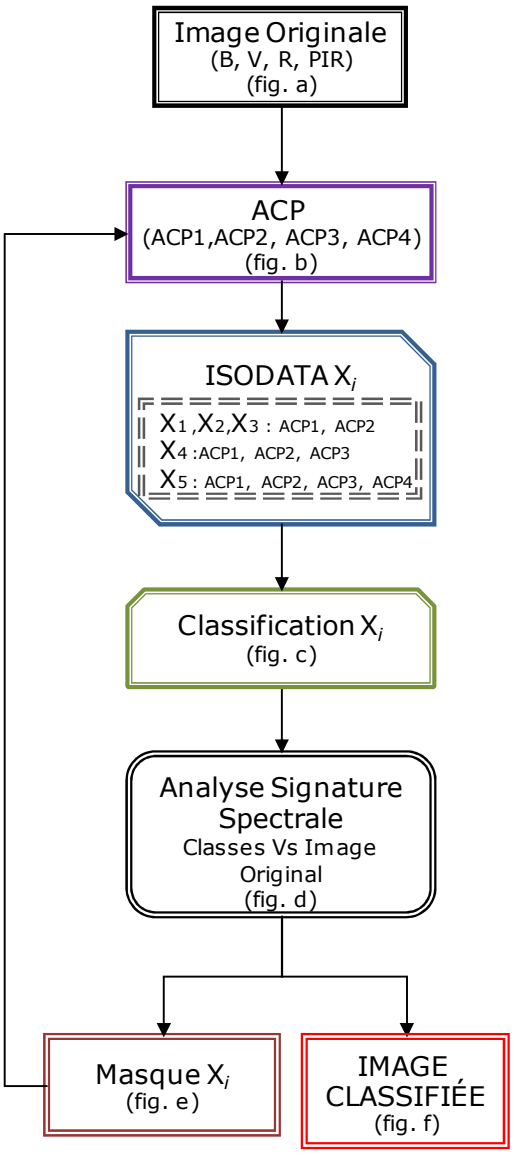
Tableau 3-1 : Pourcentage de variance expliqué lors de l'analyse en composantes principales

	ACP1	**ACP2**	**ACP3**	**ACP4**
GeoEye	84,9%	13,7%	1,2%	0,2%
Orthophoto	96,7%	2,5%	0,6%	0,2%

[32] « Le masquage consiste à cacher une partie de l'image et à conserver intacte l'autre partie » (Girard et Girard, 1999)

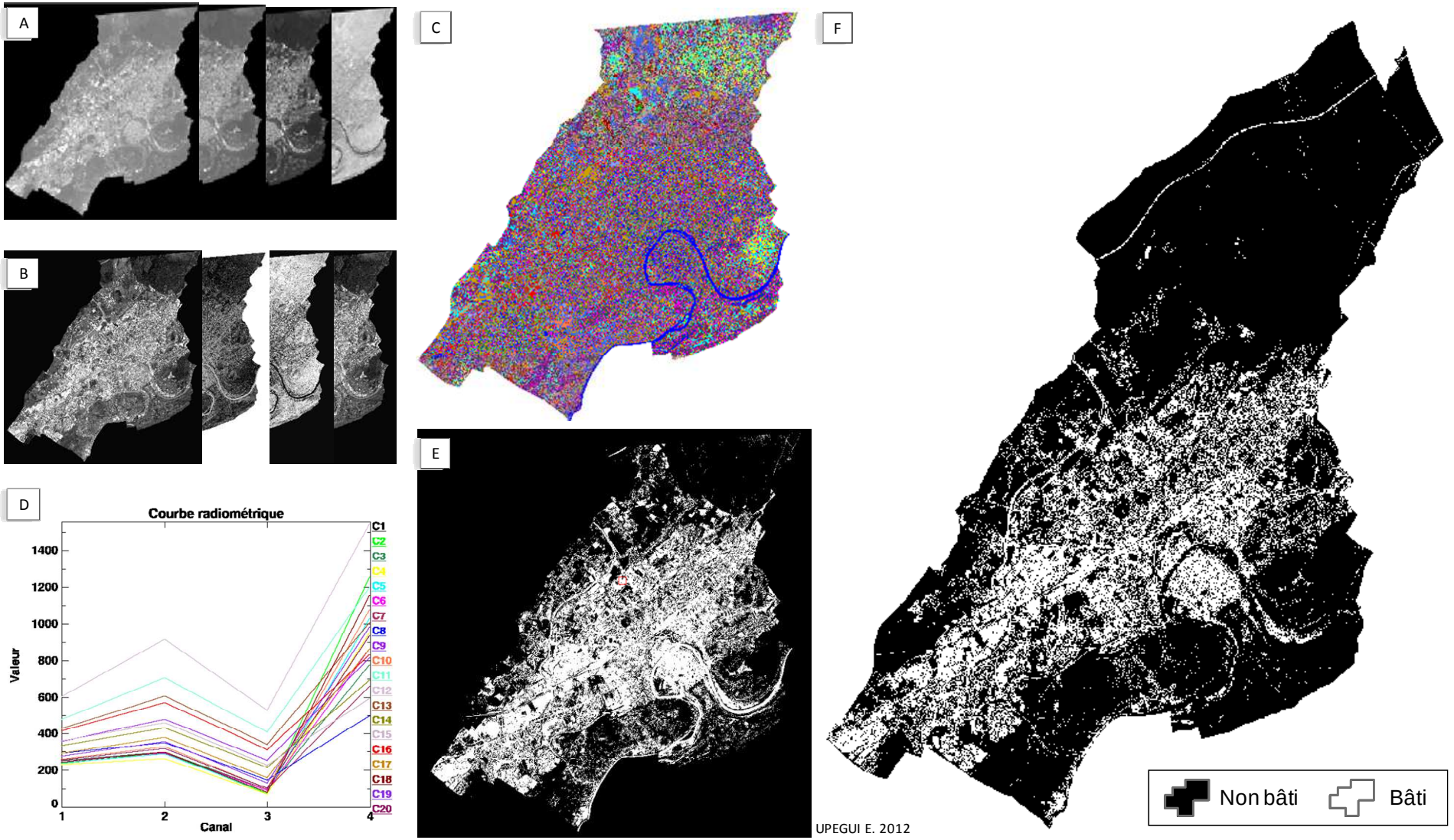
[33] « Ensemble de caractéristiques conditionnant l'interaction du rayonnement électromagnétique avec la matière, nécessaires et suffisantes pour identifier une surface déterminée », la réflectance étant l'une de ces caractéristiques les plus employées (CILF, 1997)

Figure 3-4 : Algorithme de la classification ISODATA



Sur l'image 5f, observons que la végétation a été globalement bien classée dans la categorie « Non bâti », tandis que les voies (routes et rues) ont été incluses dans la catégorie « bâti ».

Figure 3-5 : Illustration de différentes étapes de l'algorithme de la classification ISODATA pour l'extraction des bâtiments

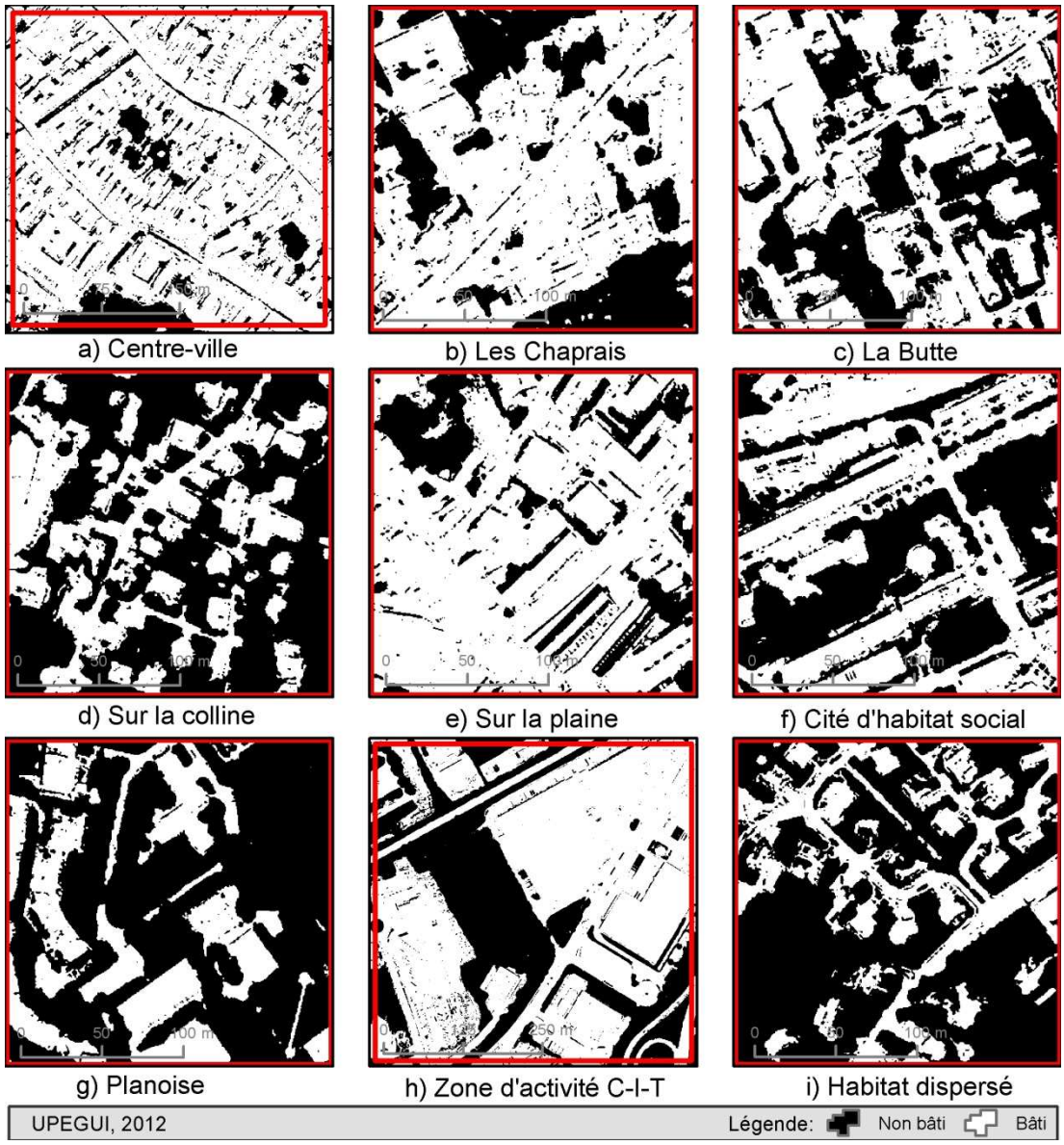


UPEGUI E. 2012

En analysant en détail les résultats de la classification ISODATA dans chacune des typologies des quartiers (fig. 3-6), nous remarquons que dans le centre-ville (fig. 3-6a) les pâtés de maisons ont été reliés par les routes. Tandis que les ombres des accidents de toiture, et des bâtiments eux-mêmes qui se reflètent sur les bâtiments et sur les routes, constituent des trous à l'intérieur des pâtés de maisons. Les cours intérieures végétalisées reproduisent ce même effet. Ces deux aspects, la connexion entre « bâti » et « non bâti » du fait des voies, et l'effet des ombres qui créent des trous sur le « bâti », se maintiennent quelque que soit la typologie des quartiers. Aux Chaprais (fig. 3-6b), nous apercevons de légers traits « non bâti » sur la route et sur l'accotement, traits qui ont été produits par les contrastes dus à la différence de matériaux entre la chaussée et le trottoir, ainsi que le marquage de la chaussée. Dans les cités d'habitat social (fig. 3-6f), et sur la plaine (fig. 3-6e), en raison de la hauteur des bâtiments et des conditions de prise de vue de l'image (altitude du soleil et du satellite), les façades sont bien éclairées, permettant la séparation entre bâtiment et route. Tandis qu'à Planoise (fig. 3-6g), étant donné l'orientation des bâtiments, l'éclairement des façades ne permet pas de détacher le bâtiment de la route, avec, de plus les toitures en terrasses qui se mêlent avec les routes. Quant à l'habitat dispersé (fig. 3-6i) et sur la colline (fig. 3-6d), les cours intérieures et les piscines ont été classifiées comme « bâti ». Enfin, dans la zone d'activité C-I-T (fig. 3-6h), les parkings n'ont pas pu être détachés des bâtiments, même si les voitures produisent un effet « poivre et sel » sur les parkings.

En définitive, les éléments qui ont permis de détacher les « plages de bâtiments » accolées aux routes, ont été par ordre d'importance : la couverture végétale (forêt, prés, et arbres isolés) ; l'ombre des bâtiments eux-mêmes, mais aussi des accidents de toitures ; le contraste entre matériaux différents (incluant celui-ci des voitures) ; et l'orientation des façades et leur éclairement.

Figure 3-6 : Détails des échantillons classifiés par ISODATA



3.1.2 Classification hiérarchisée : une méthode dirigée

La classification par « arbres de décision », ou « hiérarchisée » est la deuxième méthode que nous avons utilisée pour l'extraction des bâtiments. Le choix de cette méthode repose sur deux critères. Le premier est, une nouvelle fois, la reproductibilité et l'exportabilité de la méthode dans une optique de santé publique. Cette classification hiérarchisée permet de produire des procédures de classification compréhensibles par l'utilisateur, et d'application facile parce qu'elle consiste en une série de tests successifs partitionnant l'espace des données en sous-régions homogènes du point de vue des classes. Le second critère prend en compte les données dont nous disposons pour mener à terme cette recherche, en particulier les données LiDAR. En effet, l'un des avantages de cette approche hiérarchisée pour classifier les images repose sur la possibilité d'inclure à chaque niveau de décision différentes sources de données, différents jeux de caractéristiques, et même différents algorithmes.

La classification hiérarchisée est un type de classification dirigée « non paramétrique » (Richards et Jia, 2006), c'est-à-dire sans hypothèse sur la distribution des données. Ainsi, l'espace des données à classifier est segmenté d'une manière arborescente, avec des arbres comprenant : nœuds, branches et feuilles (fig. 3-7). Chaque nœud interne teste un attribut ; chaque branche correspond à une valeur d'attribut (à savoir « oui » ou « non » pour les arbres binaires - fig. 3-8b et fig. 3-8c) ; et chaque feuille correspond à une classe « unique ». Toutefois, il peut y avoir des chevauchements entre classes (fig. 3-8a et fig. 3-8c). Cette classification repose sur des règles de décision à différents niveaux (nœuds) où un pixel est étiqueté comme appartenant à une des classes disponibles ; sinon, il est laissé non classifié à ce niveau, pour être étiqueté à un autre niveau. Le critère d'appartenance est défini par un seuil lors de la segmentation ; une des méthodes pour déterminer ce seuil s'appuie sur les histogrammes[34] des images. Selon Girard et Girard (1989), « un histogramme à 2 modes indique que l'on est en présence d'au moins 2 populations (classes, types d'objets) différentes » (fig. 3-9). Par ailleurs, il est recommandé de réduire au minimum le nombre des caractéristiques utilisées dans une

[34] « Pour un canal, l'histogramme représente la fréquence de chacune des valeurs possibles dans ce canal » (Girard et Girard, 1989)

décision, ce qui réduit la durée du temps de calcul et améliore l'exactitude des échantillons d'apprentissage (Richards et Jia, 2006).

Figure 3-7 : Schéma d'un arbre de décision

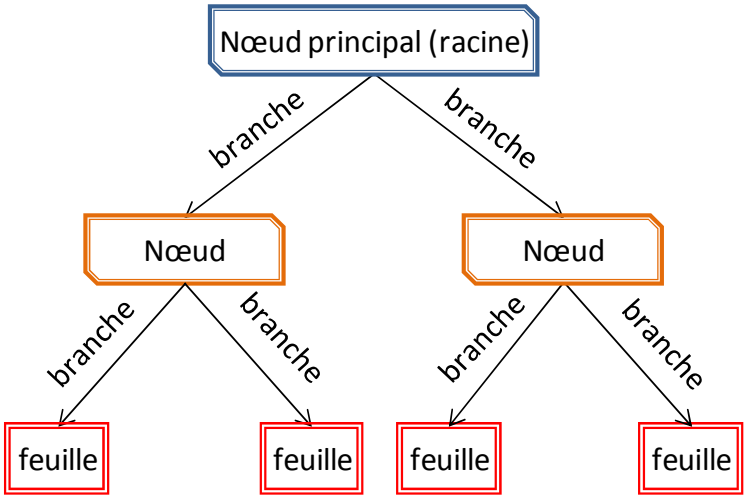


Figure 3-8 : Exemples d'arbres de décision, d'après Richards et Jia (2006). (a) arbre de décision général, (b) arbre de décision binaire avec superposition de classes, (c) arbre de décision binaire sans superposition de classes

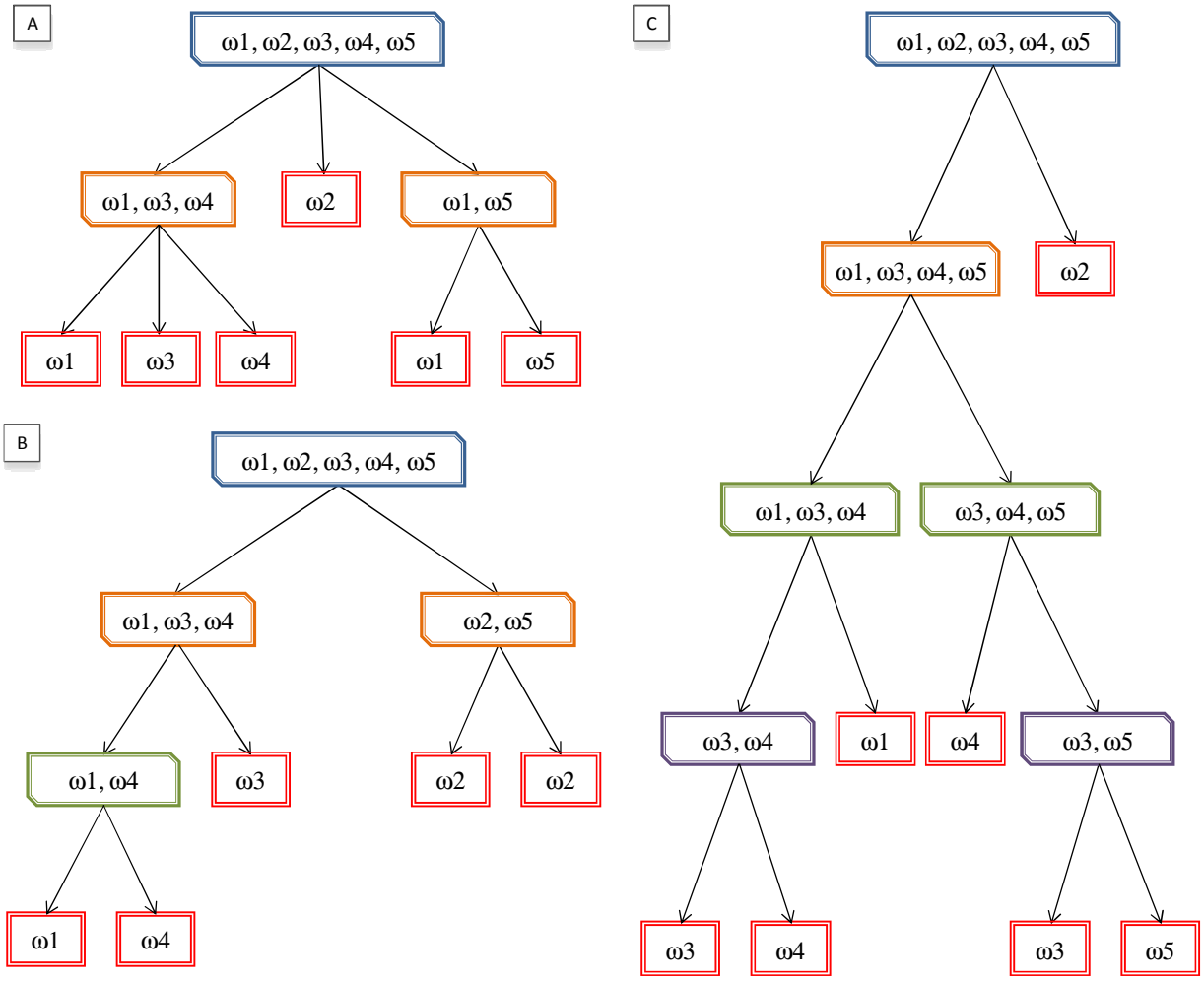
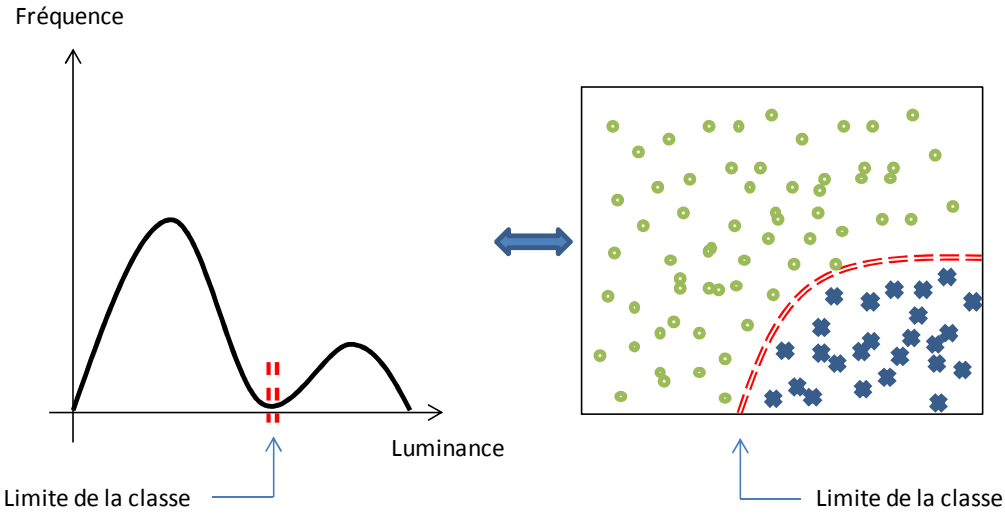


Figure 3-9 : Histogramme bimodal inspiré de Girard et Girard (1989), et sa représentation dans l'espace



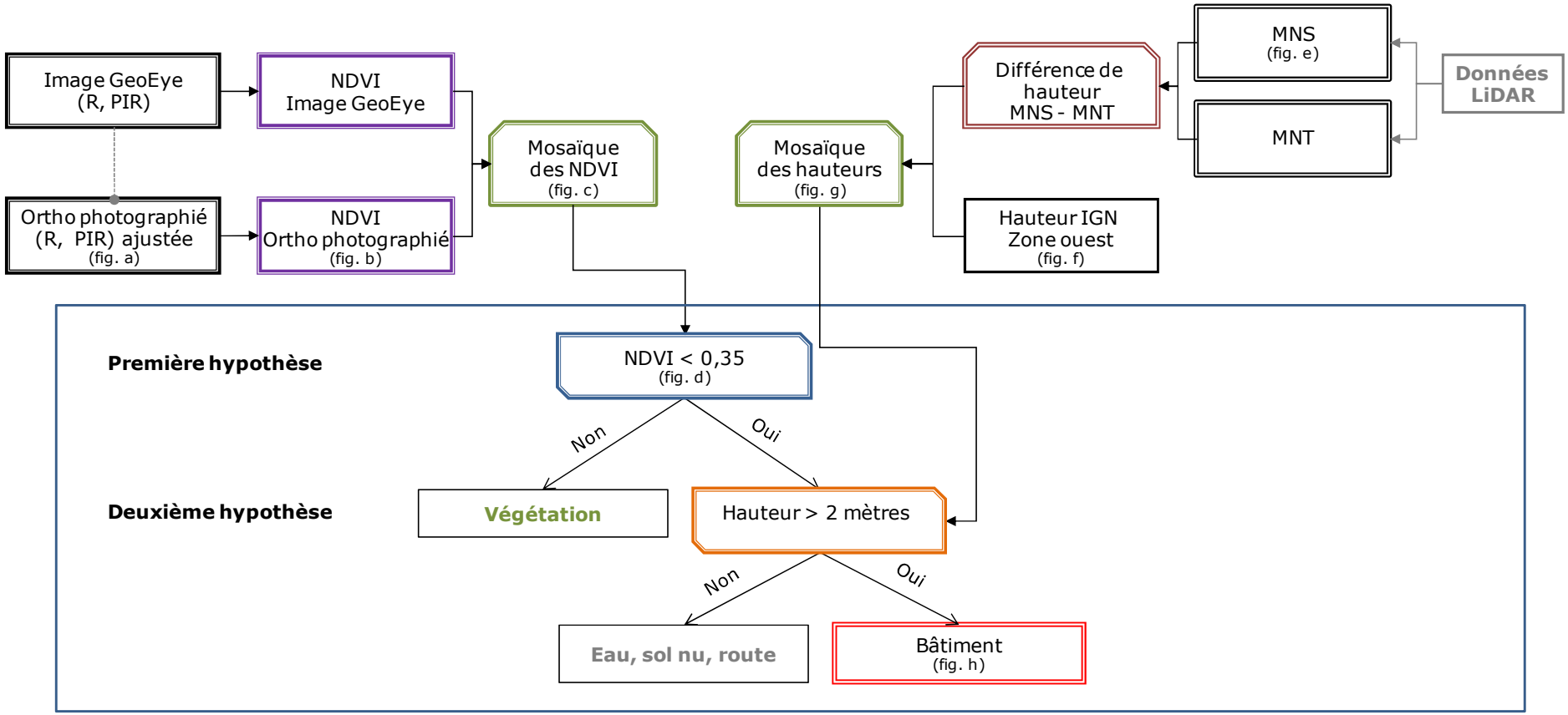
L'application de cette méthode à notre zone d'étude (fig. 3-10), repose sur deux hypothèses qui correspondent d'abord à un critère biophysique : l'absence de couverture végétale ; et ensuite à un critère géométrique : la hauteur supérieur à 2m. Pour obéir au critère biophysique, nous avons eu recours à l'indice de végétation, notamment à l'indice de la différence de végétation normalisé (NDVI), car : « l'indice de végétation permet de séparer plus ou moins aisément les zones de bâti et celles de végétation » (Weber, 1995). Les « indices de végétation » permettent l'identification des couverts végétaux, en se fondant, d'une part, sur la luminance provenant des objets au sol, et d'autre part, sur les différences existantes entre le comportement spectral des végétaux et celui des sols (Girard et Girard, 1999). Ces différences sont particulièrement marquées dans les canaux : rouge (R) et proche infrarouge (PIR). Le calcul du NVDI $\left(\frac{PIR-R}{PIR+R}\right)$ (Tucker et Sellers, 1986) peut se faire sur les comptes numériques des images (sans les transformer en valeur de réflectance ou de luminance), et ce dans le cas où l'on ne cherche qu'une valeur « relative » de la présence de végétation (Chuvieco, 1996). En ce qui concerne le critère géométrique, nous avons calculé la hauteur des bâtiments à travers le MNT et le MNS (issues de données LiDAR), en soustrayant la hauteur au sol (MNT) de la hauteur au toit (MNS) (Dong *et al.*, 2011 ; Lu *et al.*, 2011 ; Wu *et al.*, 2008).

En pratique, pour l'extraction des caractéristiques (première étape de la classification), nous avons préparé les données afin d'effectuer des tests dans chacun des deux nœuds de notre arbre (nos hypothèses). D'une part, nous avons ajusté l'histogramme (canal à canal, le canal R et PIR) de l'orthophotographie aérienne à celui de

l'image GeoEye (fig. 3-11a), en nous appuyant sur l'algorithme « histogram matching[35] » (Richards et Jia, 2006), en raison de la différence de ces caractéristiques spectrales (cf. chapitre 2 : tab. 2-1 et tab. 2-3). Puis le NDVI a été calculé tant pour l'orthophoto (fig. 3-11b) que pour l'image GeoEye ; et ensuite une mosaïque a été réalisée afin d'avoir une seule image de la végétation (fig. 3-11c). D'autre part, nous avons calculé la différence de hauteur (fig. 3-11g), avec la formule $Hauteur = MNS - MNT$ (le MNS est illustré dans la fig. 3-11e), pour la totalité de la couverture des données LiDAR. Cependant, ces données ne recouvrent pas toute la zone d'étude (cf. chapitre 2 : fig. 2-18). C'est pour cette raison que nous avons utilisé la BDTopo® (données de référence) afin d'obtenir les hauteurs des bâtiments dans la partie ouest de la ville (fig. 3-11f). Puis une mosaïque a été faite pour avoir une seule image avec les différences de hauteur de la totalité de notre zone d'étude. En ce qui concerne l'apprentissage des pixels (deuxième étape de la classification), dans l'hypothèse d'une absence de couverture végétale nous avons déterminé le seuil à 0,35 en raison de l'inversion de l'histogramme (fig. 3-11d) du NDVI (Girard et Girard, 1989). Quant à la hauteur, pour en identifier le seuil, nous nous sommes inspirée des travaux de Dong et de son équipe (2011) qui ont proposé un seuil de 2,2m. Toutefois nous avons arrondi cette valeur à 2m, en raison de la précision verticale des données altimétriques (à savoir 15cm, selon le fournisseur - SAF, 2009). Enfin, pour l'étiquetage (troisième étape de la classification) en raison du type de stratégie, à savoir « dirigée », il s'est fait « automatiquement », en suivant la règle de décision déduite de nos hypothèses : « SI le *NDVI* est inférieur à [0,35] ET la *Hauteur* est supérieure à [2m], ALORS le pixel correspond à un « Bâtiment » (fig. 3-11h) ». Ainsi, nous observons sur la figure 3 - 11h, de manière globale, que végétation et routes ont été bien classées dans la catégorie « non bâti ».

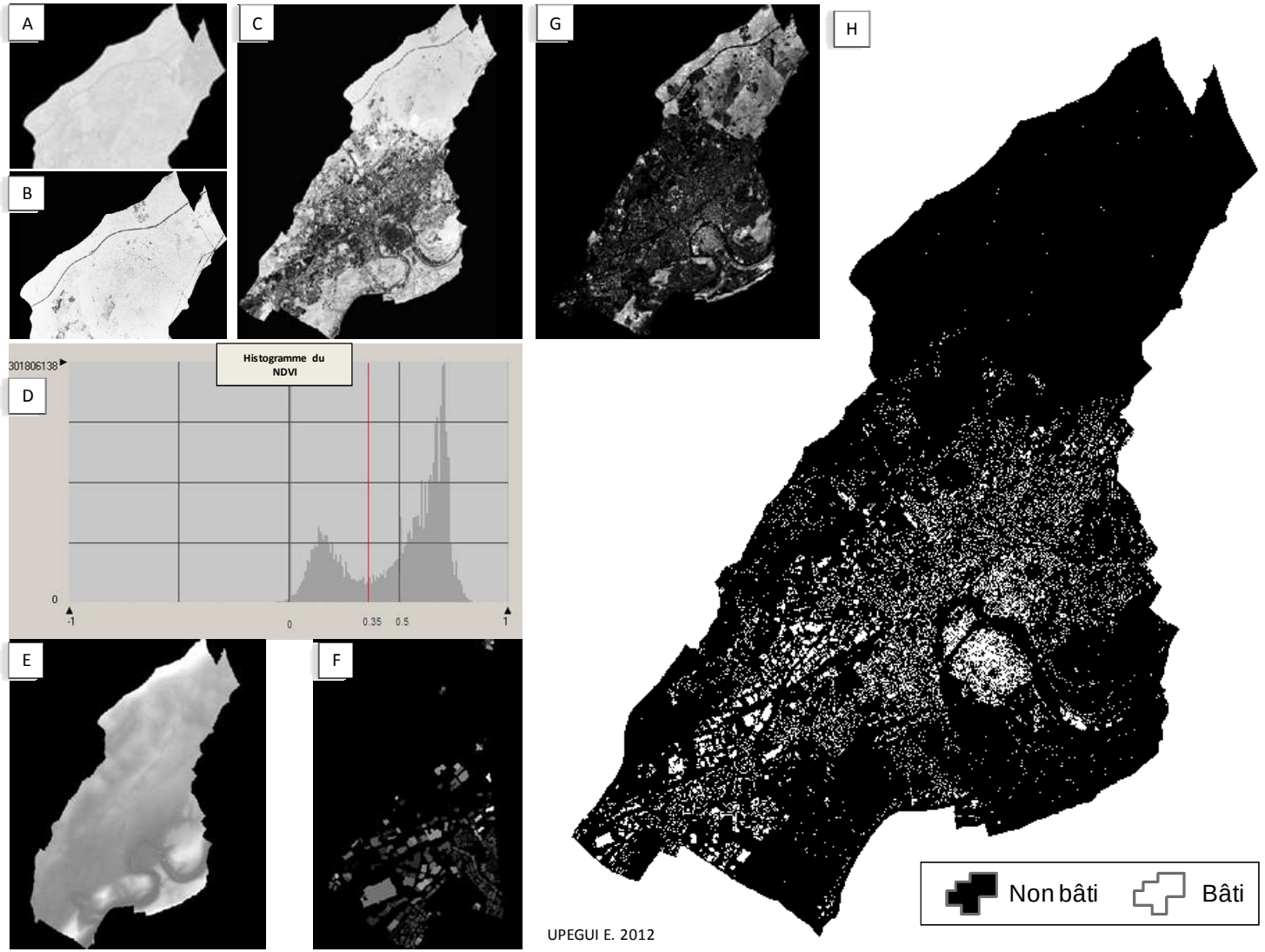
[35] Cet algorithme détermine mathématiquement une table de coloration (« look up table », en anglais) qui s'emploie pour transformer l'histogramme de distribution de fréquences d'une image donnée en celui d'une autre, afin qu'elles se ressemblent (Richards et Jia, 2006).

Figure 3-10 : Algorithme de la classification hiérarchisée



UPEGUI E. 2012

Figure 3-11 : Illustration de différentes étapes de l'algorithme de la classification hiérarchisée pour l'extraction des bâtiments

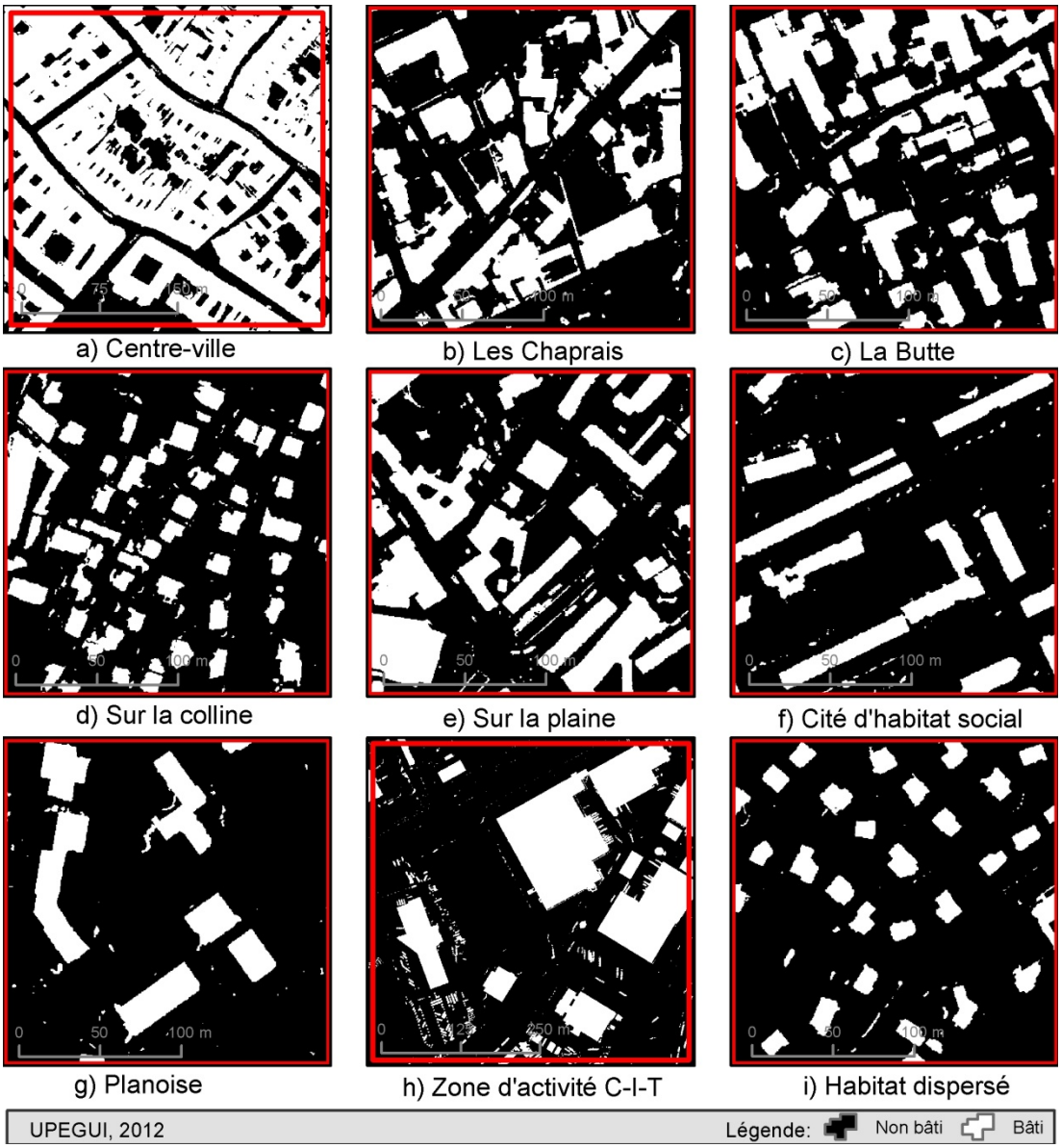


Après analyse détaillée des résultats de la classification hiérarchisée dans chacune des typologies des quartiers (fig. 3-12), nous remarquons que dans le centre-ville (fig. 3-12a) les pâtés de maisons ont été clairement définis et complètement détachés des routes. Les cours intérieures, végétalisées ou non, ont été bien classées dans la catégorie « non bâti ». Le détachement des bâtiments des routes se maintient dans les différentes typologies des quartiers. Aux Chaprais (fig. 3-12b), nous apercevons de légers traits « bâti » correspondant aux murs divisant les terrains. Dans les cités d'habitat social (fig. 3-12f), sur la plaine (fig. 3-12e), et à Planoise (fig. 3-12g) les immeubles ont été bien classés ; les toitures en terrasses n'ont, dans cette classification, posé aucun problème. Toutefois, on peut remarquer deux limites à cette classification. La première correspond à l'ombre des arbres isolés, qui n'a pas été supprimée à travers le critère biophysique. La deuxième limite consiste en la différence entre les dates ainsi que les conditions de prise de vue des données altimétriques et optiques, se traduisant par la présence d'éléments que nous ne voyons pas sur l'image (GeoEye ou orthophoto) mais qui ont une hauteur supérieure à 2 m. C'est pour cette raison que certains éléments, qui ont une forte chance d'être des bus ou des camions (apparaissant sur les routes, et ayant une taille et une forme précises), sont classés comme « bâti ». Quant à l'habitat dispersé (fig. 3-12i), sur la colline (fig. 3-12d) et encore à La Butte (fig. 3-12c), les cours intérieures ont été bien détachées des bâtiments, et donc ont été classifiées comme « non bâti ». Enfin, dans la zone d'activité C-I-T (fig. 3-12h), les parkings ont été bien détachés des bâtiments, même si, sur les parkings, les voitures apparaissent comme éléments classées comme « bâti ».

En définitive, ces deux critères, biophysique et géométrique, ont permis de séparer aisément les « bâtiments » des routes, des cours intérieures, et en général d'autres zones imperméables[36]. En outre, les accidents de toiture ou le contraste entre matériaux de construction différents ne jouent pas un rôle majeur dans cette classification.

[36] Une superficie dure qui empêche ou qui retarde la pénétration de l'eau dans les sols comme dans des conditions normales (avant l'urbanisation), ou qui laisse ruisseler l'eau sur la surface en plus grande quantité qu'avant le développement (Agence de Protection de l'Environnement des Etats-Unis (siglé « EPA », en anglais) http://www.epa.gov/owow/NPS/MMGI/Chapter4/ch4-8.html (consulté le 13 mais 2011).

Figure 3-12 : Détails des échantillons classifiés par la classification hiérarchisée



3.2 Classifications fondées sur l'objet

Cette stratégie de classification, *a priori* est plus adéquate pour les données à très haute résolution spatiale (Blaschke, 2010 ; Puissant *et al.*, 2004), reste encore un « défi » pour les télédétecteurs.

3.2.1 Segmentation de l'image par morphologie mathématique

L'approche « orientée-objets » s'emploie de plus en plus dans des applications de télédétection en raison de la mise à disposition des données à THRS (Weng, 2010). Cette approche s'applique à des régions préalablement segmentées dans une image ; la segmentation des images fusionne les pixels en « objets », permettant ainsi la mise en œuvre d'une classification à partir des « objets » au lieu des « pixels individuels » (Weng, 2010). Néanmoins, la segmentation doit produire des régions spatialement continues, homogènes, mais disjointes (Blaschke *et al.*, 2005). L'approche « orientée-objets » se révèle être un outil puissant pour l'analyse d'images (Blaschke *et al.*, 2005 ; Matsuyama et Hwang, 1990 ; Forestier *et al.*, 2012). C'est pour cela que « la segmentation est sans doute la tâche qui, en analyse d'images, mobilise le plus d'effort » (Beucher, 1990 ; Vachier, 1995). A l'heure actuelle, il existe différents logiciels commerciaux dédiés à la classification fondée sur l'objet, tels *Defiens* (auparavant appelé *e-cognition)* ou encore *Feature Extraction* (progiciel d'ENVI), qui tendent à faciliter la segmentation. Néanmoins, cette segmentation reste intuitive (sans hypothèse de départ) ; subjective (opérateur dépendant) ; et repose sur la mise au point des objets en zoomant de façon croissante/décroissante, à partir du pixel (Chen *et al.,* 2009 ; Jacquin *et al.,* 2008 *;* Jahjah et Ulivieri, 2010 ; Lu *et al.,* 2010, Van Der Sande *et al.,* 2003 ; entre autres). Bien que les résultats de ces classifications soient satisfaisants, la reproductibilité de cette approche reste limitée suite aux contraintes exposées ci-dessus.

Les différentes méthodes de segmentation peuvent être regroupées en trois approches : « par pixel », « par frontière » et « par région » (Blaschke *et al.*, 2005). Selon Blaschke et son équipe, les méthodes « par pixel » incluent « le seuillage d'images » (*images thresholding*, en anglais), et la segmentation « par caractéristiques spatiales ».

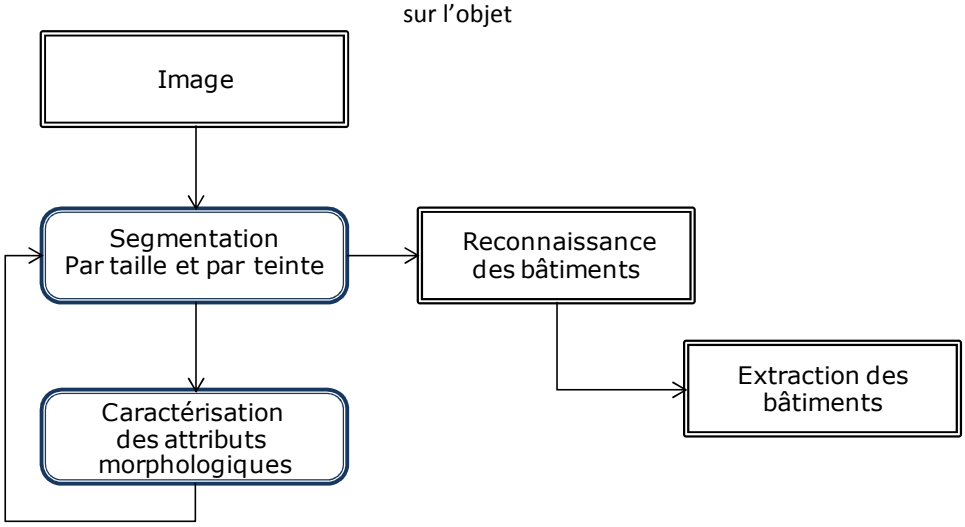
Les méthodes « par frontière » comprennent les « filtrages » et les « gradients » des images qui mettent en évidence les contours des objets ; la « ligne de partage des eaux » (Beucher et Lantuejoul, 1979) fait partie de ce type de méthode. Les méthodes « par région » englobent différentes techniques comme : « les régions croissantes », « la fusion des régions », et « la désintégration des régions ».

Au-delà, la segmentation doit répondre à une hypothèse de reconnaissance de l'objet que l'on cherche, afin de garantir la reproductibilité de la méthode. Nous avons donc émis l'hypothèse que l'on peut extraire les bâtiments -autrement dit, l'objet que l'on cherche- en deux temps (fig. 3-13). Dans un premier temps, on peut reconnaitre les bâtiments par leur teinte et par leur taille ; dans un deuxième temps, on peut les extraire grâce aux attributs morphologiques des objets segmentés. Cela rejoint les deux grandes tâches nécessaires préalables à la compréhension des images (Matsuyama et Hwang, 1990). Le critère de teinte permettrait de récupérer les différents pans d'une toiture, qui peuvent varier de couleur, suivant l'éclairement du soleil. Le critère de taille permettrait d'identifier les bâtiments de surface différente (par exemple : petite, moyenne et grande), sans entrer dans le détail du type de tissu urbain (Puissant et Weber, 2004). Enfin, les critères morphologiques pallieraient l'absence de données altimétriques, et permettraient la différenciation des bâtiments par rapport aux autres objets (zones imperméables) (Nagao *et al.*, 1978). Pour segmenter l'image et obtenir les objets, nous nous sommes appuyée sur des principes de morphologie mathématique (MM) (Serra, 1982). D'ailleurs, Jahjah et Ulivieri, en 2010, et plus récemment, en 2012, Forestier et son équipe ont démontré les avantages à extraire les objets par MM au lieu de les extraire en utilisant *Defiens* ou *Feature Extraction*

La définition de MM, selon Bloch (2008), est la suivante :

La morphologie mathématique est une théorie essentiellement non linéaire, utilisée en particulier en analyse d'images, dont le but est l'étude des objets en fonction de leur forme, de leur taille, des relations avec leur voisinage (en particulier topologiques), de leur texture, et de leurs niveaux de gris ou de leur couleur.

Figure 3-13 : Organigramme des étapes générales pour extraire des bâtiments avec une approche fondée sur l'objet



Les méthodes de la MM permettent d'extraire un objet donné sur une image : elles le mettent en évidence et ensuite le déconnectent, en utilisant uniquement l'information à l'intérieur de la même image. Pour aboutir à cet objectif, l'analyse morphologique de X s'effectue par l'intermédiaire de transformations ensemblistes[37] ψ à l'aide d'un élément structurant (ES) de géométrie simple (Mering, 1990). Dans la MM, l'analyse d'image ne répond pas à une classe de phénomènes particuliers mais elle s'intéresse à l'aspect « spatial » des objets étudiés (Serra, 1982). Ainsi, la transformation la plus adéquate pour extraire un objet dépend des contraintes spécifiques pour ce cas particulier. Dans notre cas, pour le critère de la teinte, les opérations suivies correspondent aux « ouvertures par reconstruction géodésique ($\varphi(f)$) » et aux « fermetures par reconstruction géodésique ($\gamma(f)$) ». En effet, ces opérations éliminent les petites structures en préservant les contours des structures plus grandes que la taille de l'ES (Coster et Chermant, 1989). Quant au critère de la taille, nous avons eu recours à l'analyse granulométrique (AG) des images (Matheron, 1975), car cette approche avait été déjà utilisée pour cartographier les surface bâties à partir des images satellite à haute résolution spatiale (Mering et Chopin, 2002 ; Chopin et Mering, 2004). L'AG est l'étude de la taille des objets fondée sur le principe du tamisage, ce qui permet la sélection des objets par un ensemble de tamis de différentes tailles. Cet ensemble d'images représente le « profil granulométrique » (ou morphologique) de l'image originale.

[37] Les définitions des opérations de MM se trouvent dans l'annexe 1.

Ainsi, la segmentation a été faite avec une approche « par région », en particulier avec la méthode de la « dérivée du profil morphologique » (DPM) proposée par Pesaresi et Benediktsson (2001). Cette approche, DPM, peut être considérée comme analogue à la technique des « régions croissantes ». Cependant, au contraire de cette dernière, qui utilise les propriétés statistiques locales, le DPM emploie la similitude des pixels, fondée sur les caractéristiques morphologiques des composantes connexes sur les images (Pesaresi et Benediktsson, 2001). La DPM est un résiduel morphologique entre la fonction de l'image originale en teinte de gris (*f*) et leur granulométrie, ce qui permet d'analyser la variation entre niveaux du profil morphologique. Cette approche nous permet donc d'intégrer nos hypothèses de départ (taille, teinte), ainsi que les opérations que nous avons envisagé d'utiliser (granulométrie, ouverture/fermeture par reconstruction géodésique). A la différence du travail de Pesaresi et Benediktsson (2001), nous utilisons les transformations par reconstruction géodésique au lieu d'ouvertures et de fermetures algébriques. Les DPM sont définies ainsi :

$$DPM\ \varphi(f)\,\mathrm{ES_i} = f - \varphi(f)\,\mathrm{ES_i}$$
$$DPM\ \gamma(f)\,\mathrm{ES_i} = \gamma(f)\,\mathrm{ES_i} - f$$

Où (*f*) est la fonction de l'image originale en teinte de gris, ES_i est l'ES de taille *i*, $\varphi(f)$ est l'ouverture par reconstruction géodésique, et $\gamma(f)$ correspond à la fermeture par reconstruction géodésique.

Du point de vue pratique, une segmentation multi-niveau (fig. 3-14) a été appliquée aux images en teinte de gris, c'est-à-dire sur l'image GeoEye panchromatique, ainsi qu'à la première composante de l'ACP de l'orthophoto (à défaut de canal panchromatique). La végétation et le cours d'eau sont masqués au préalable : pour la végétation nous avons retenu le même critère que pour la classification hiérarchique (NDVI < 0,35), et pour l'eau nous avons eu recours aux opérations de morphologie mathématique (seuillage bas et ensuite une ouverture par reconstruction avec un ES de taille 50). La forme « disque » a été choisie comme ES : le disque, en effet, ne privilégie aucune direction (isotrope), permettant ainsi d'identifier les bâtiments sans tenir compte de leur orientation. La taille de l'ES a varié entre 5 et 100, permettant ainsi de créer trois niveaux d'analyses (NA) pour chacune des deux transformations ($\varphi(f)$ (fig. 3-15), $\gamma(f)$ (fig. 3-16)). Le premier niveau -

« petit bâtiment »- a fait varier la taille de l'ES entre 5 et 25 avec un pas de 5. Dans le deuxième niveau -« bâtiment moyen »- la taille de l'ES a varié entre 25 et 60, avec un pas de 10 à partir de la taille 30. Le dernier niveau d'analyse -« grand bâtiment »- correspond aux variations de la taille de l'ES entre 60 et 100 avec un pas de 10. Ensuite, trois classifications non dirigées[38] par masques emboîtés ont été réalisées sur chacun des six niveaux d'analyses (trois pour chaque type de DPM). Au total, 18 classifications ont été opérées sur l'ensemble de l'image GeoEye, alors qu'une seule classification a été faite sur l'ortho-photo. La figure 3 - 17 présente un exemple de la première classification, faite sur le niveau d'analyse « moyen », découlant des DPMφ(f). L'analyse de la courbe granulométrique (fig. 3-18) indique qu'il existe 5 classes de bâtiments potentiels (BP) (fig. 3-19), à savoir : la classe 5 qui présente un pic dans la granulométrie de taille 60 ; les classes 6 et 9 avec un pic dans la granulométrie de taille 50 ; la classe 7 qui représente la granulométrie de taille 25 ; et la classe 8 qui montre un pic dans la granulométrie de taille 40. Ainsi, lors de l'application de cette méthode, nous avons identifié 74[39] couches de BP (2 entre elles sur l'ortho-photo), distribuées de la manière suivante. En ce qui concerne l'image GeoEye, pour le niveau d'analyse « petit », nous avons identifié 14 classes de BP par la DPM φ(f), et 7 classes de BP par la DPM γ(f). Pour le niveau d'analyse « moyen », 17 classes de BP ont été identifiées par DPM φ(f) tandis que par la DPM γ(f) nous avons répertorié 12 classes de BP. Enfin pour le niveau d'analyse « grand », 16 classes de BP ont été identifiées par chacune des DPM (φ(f), γ(f)). S'agissant de l'orthophoto, nous avons identifié seulement 2 classes de BP dans la DPM φ(f), notamment dans le niveau d'analyses « petit ».

[38] K-mens a été l'algorithme utilisé ; le nombre de classes et le nombre d'itérations ont été fixés à 9, et le seuil de convergence a été fixé à 95%.
[39] Le détail de ces couches se trouve dans l'annexe 2.

Figure 3-14 : Schéma général de la segmentation multi-niveaux

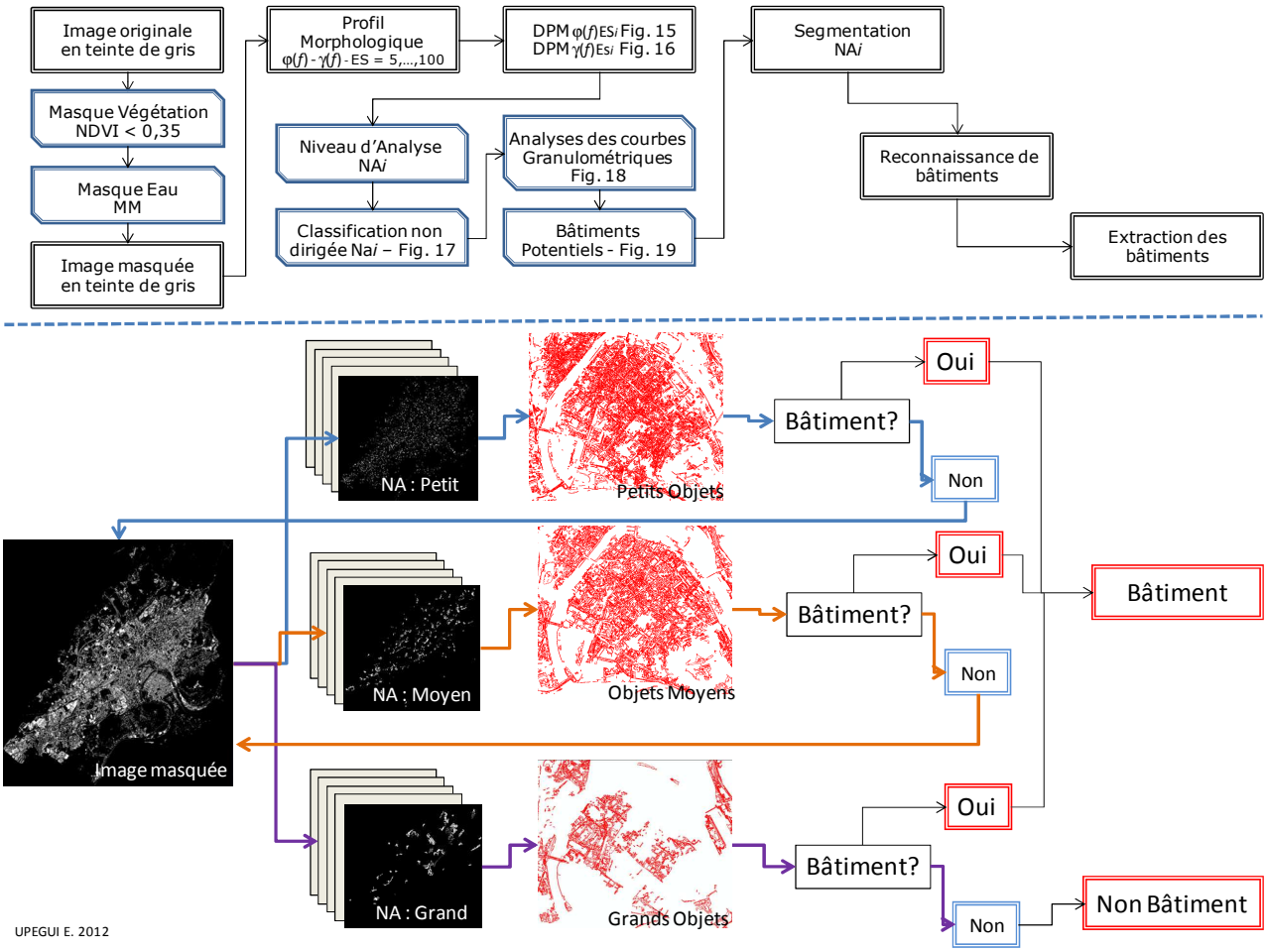


Figure 3-15 : Les trois niveaux d'analyse découlant de la DPM φ(f) sur l'image GeoEye. En haut le niveau « grand bâtiment », au milieu le niveau « bâtiment moyen », et en bas le niveau « petit bâtiment »

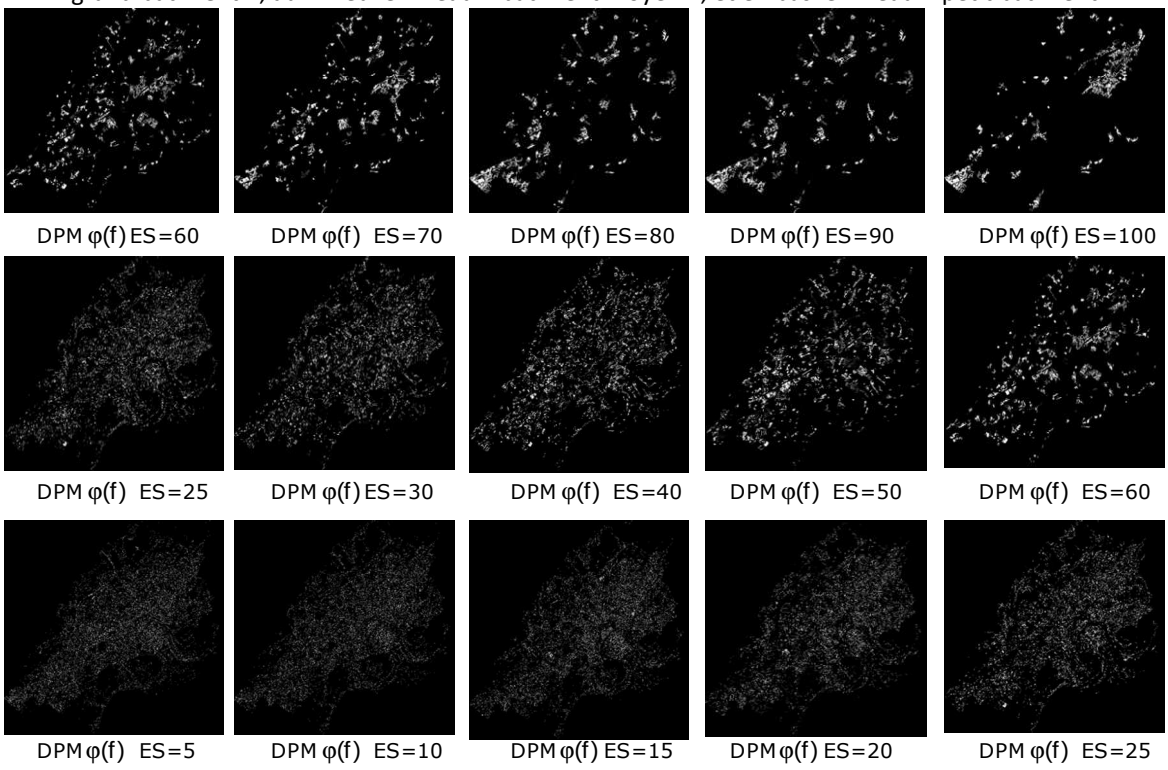


Figure 3-16 : Les trois niveaux d'analyse découlant de la DPM γ(f) sur l'image GeoEye. En haut le niveau « grand bâtiment », au milieu le niveau « bâtiment moyen », et en bas le niveau « petit bâtiment »

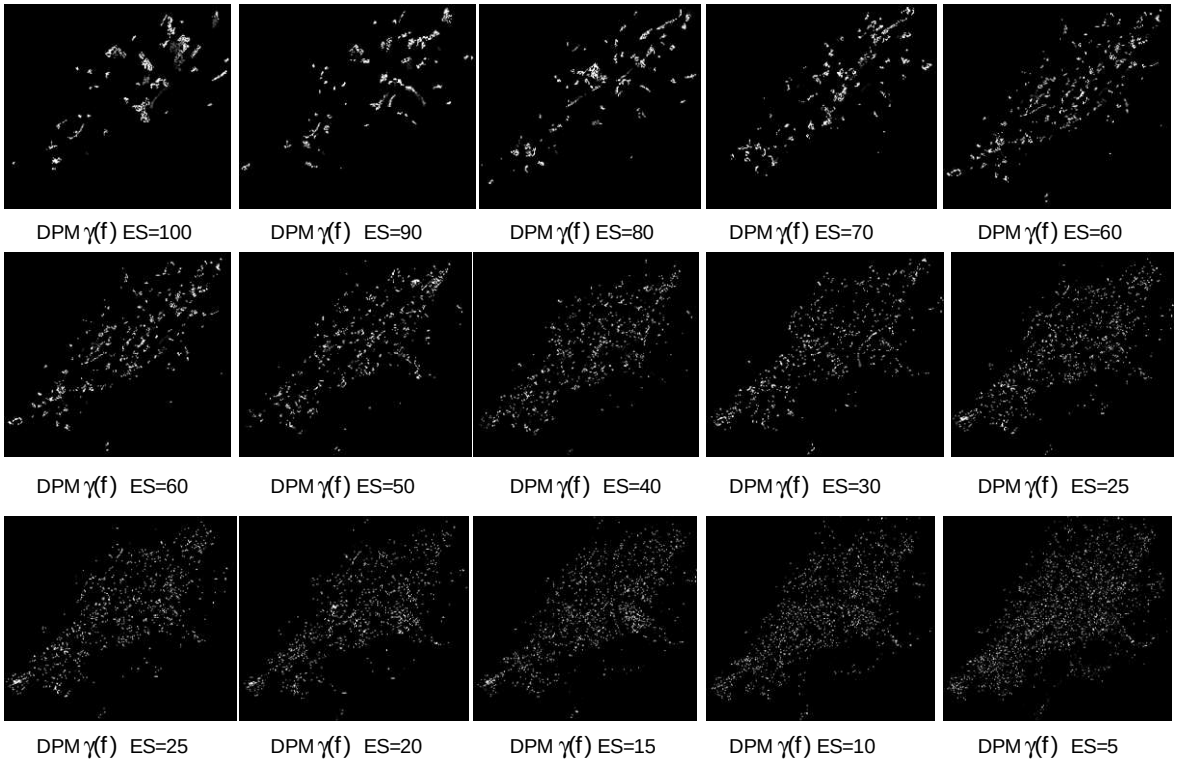


Figure 3-17 : Classification du niveau d'analyse « moyen » de la DPM φ(f)

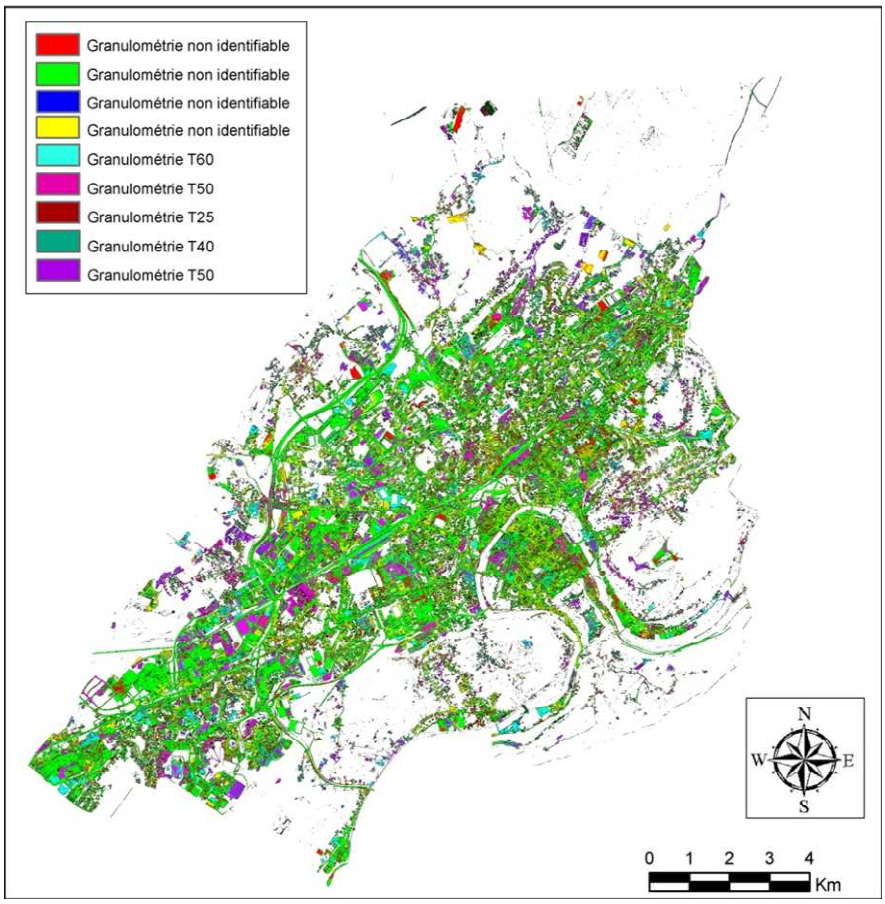


Figure 3-18 : Analyse des courbes granulométriques de la classification du niveau d'analyse « moyen » de la DPM φ(f)

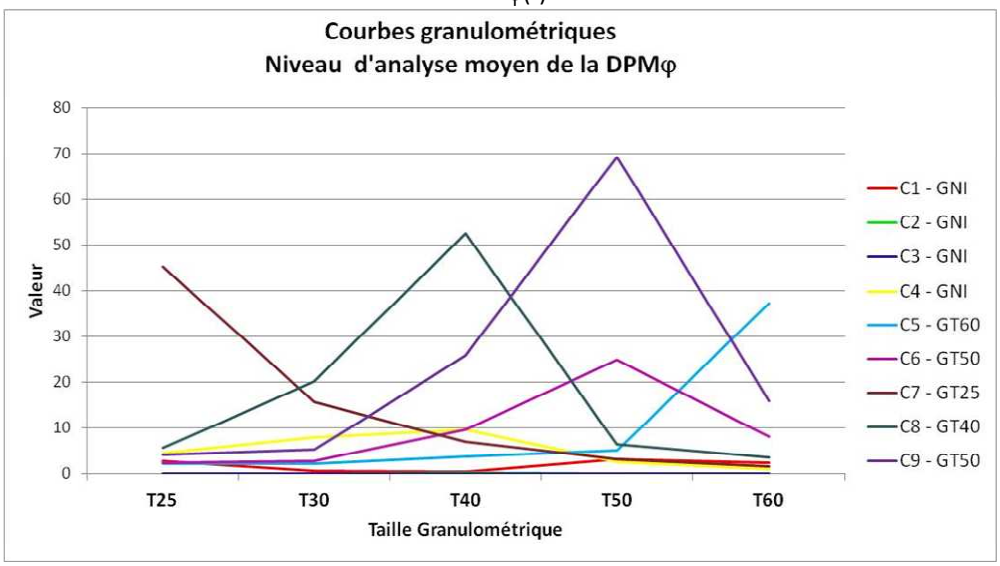
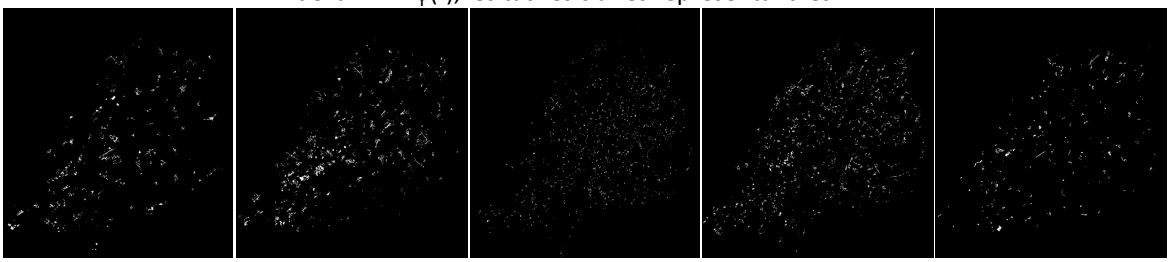


Figure 3-19 : Bâtiments potentiels identifiés dans la classification du niveau d'analyse « moyen » découlant de la DPM φ(f), les taches claires représentant les BP.



BP Classe 5 – Granulométrie T60 | BP Classe 6 – Granulométrie T50 | BP Classe 7 – Granulométrie T25 | BP Classe 8 - Granulométrie T40 | BP Classe 9 – Granulométrie T50

En définitive, pour segmenter les images de notre zone d'étude, 52 nouvelles images en teinte de gris ont été créées à partir de chacune des deux images en teinte de gris (GeoEye et orthophoto). Ces 52 images se distribuent de la manière suivante : les 13 premières images correspondent au profil morphologique (PM) de l'ouverture par reconstruction géodésique φ(f) ; les 13 images suivantes correspondent au PM de la fermeture par reconstruction géodésique γ(f) ; le groupe suivant de 13 images concerne les dérivés du profil morphologique (DPM) découlant de l'φ(f) ; et le dernier groupe de 13 images est consacré aux DPMγ(f). Ensuite, 22 classifications (18 sur GeoEye et 4 sur l'orthophoto) ont été opérées sur les DPM (φ(f) et les DPM γ(f)) ; chacune des classifications a inclus seulement 5 images, chacune correspondant à un niveau précis d'analyse. Résultant de l'analyse des courbes granulométriques de ces classifications, 74 couches de bâtiments potentiels ont été identifiées, et environ 130 000 objets ont été créés. Nous avons choisi, par tirage aléatoire, un échantillon d'étude correspondant à 10% du total de ces objets, constituant ainsi un corpus de 13 097 objets à analyser.

3.2.2 Etude préliminaire de la potentialité des indicateurs morphologiques pour la discrimination bâti/non bâti

A l'heure actuelle un grand nombre d'attributs pour caractériser différents types d'éléments sont proposés par les logiciels. Néanmoins, la caractérisation des objets, notamment des bâtiments, reste encore un « test d'échec». Cela se constate par la variété et le nombre important de paramètres employés pour caractériser les objets (Chen *et al.,* 2009 ; Jacquin *et al.,* 2008 *;* Jahjah et Ulivieri, 2010 ; Lu *et al.,* 2010). Les bâtiments étant par nature des objets complexes, différents exemples mettent en évidence la difficulté de leur extraction à travers les résultats obtenus. En 2012, Forestier et son équipe ont employé un système à base de connaissances (SBC) pour classifier des images THRS. Dans cette recherche la précision de l'identification des bâtis est bien moindre (0,353 à Marseille et 0,589 à Strasbourg) que celle de l'identification de la végétation (0,976 à Marseille et 0,995 à Strasbourg) ou encore celle de l'identification des routes (0,991 à Marseille et 0,732 à Strasbourg). De plus, il faut signaler que la typologie des bâtiments, dans les deux cas (à Marseille et à Strasbourg) correspond à du pavillonnaire. Cette même typologie de bâtiment a été employée dans l'expérimentation réalisée par Nagao et son équipe (1979, 1979b), ainsi que par Benediktsson et son groupe (2003). Toutefois, ces derniers ont aussi inclus des grands immeubles tout en restant dans un habitat dispersé. Sur cette même zone d'étude, Chanussot et son équipe, en 2006, ont extrait les bâtiments par deux méthodes différentes avec une précision globale de 46,8% et 57,3% ; ils ont segmenté l'image avec la DPM.

C'est pour ces raisons que nous considérons important et pertinent d'approfondir la compréhension de cet objet « le bâtiment » dans le domaine de la reconnaissance des formes. Pour ce faire, nous étudions la potentialité des indicateurs morphologiques pour la discrimination « bâti/non bâti » avant de continuer avec l'extraction des bâtiments proprement dite. Cela a une double finalité : d'une part, réduire le nombre des variables à utiliser dans la caractérisation des objets ; et d'autre part, créer des « règles de sélections d'objets » plus déterministes (disons « moins empiriques »), comme cela se faisait jusqu'à présent.

Avec les objets créés résultant de la segmentation de l'image (première étape accomplie), nous abordons la deuxième grande étape (fig. 3-13) : la caractérisation des

attributs morphologiques, avec 28 attributs (mesures) (tab. 3-2) calculés pour chacun des objets extraits ; cela s'est fait avec le logiciel Aphelion. Ce logiciel calcule différents types de mesures en fonction du type d'objet (point, ligne, polygone), mais aussi des objets formés par les projections orthogonales de la frontière de celui-ci sur une ligne avec un angle donné (selon l'orientation de l'objet). Nous avons utilisé les mesures dites « classiques » de « dimension et forme », lesquelles incluent les mesures qui comparent l'objet de forme « inconnu » avec une forme géométrique simple, comme un cercle ou un rectangle (Lee et Sallee, 1970). De même, nous avons utilisé les mesures effectuées sur le « rectangle minimum délimité » (MBR, en anglais *Minimum Bounding Rectangle*) défini par les objets projetés orthogonalement. Les MBR ne sont pas généralement parallèles aux axes de coordonnées (X, Y) de l'image (fig. 3-20).

Figure 3-20 : Mesures sur un MBR

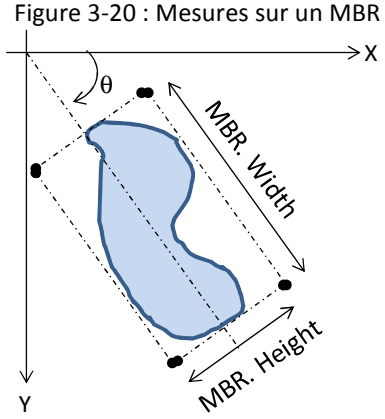


Tableau 3-2 : Attributs morphologiques calculés

AREA Surface	**CONVEXAREA** Surface de l'enveloppe convexe de l'objet
BOUNDINGRECTANGLETOPERIMETER $\frac{\text{Périmètre}}{2 * \text{Hauteur} + 2 * \text{Largeur}}$	**CONVEXPERIMETER** Périmètre de l'enveloppe convexe de l'objet
BOUNDINGRECTFILL $\frac{\text{Surface}}{\text{Hauteur} * \text{Largeur}}$	**EQUIVALENTDIAMETERP** Diamètre du cercle dont leur Périmètre = Périmètre de l'objet
CIRCULARITY $\frac{4\pi * \text{Surface}}{\text{CroftonPerimeter2}}$	**ELONGATION** $\frac{(\text{Inertie axe ppal} - \text{Inertie Axe sec.})}{(\text{Inertie axe ppal} + \text{Inertie Axe sec.})}$

COMPACTNESS $\frac{16 * \text{ Surface}}{\text{Périmètre} * 2}$	**EQUIVALENTDIAMETER** Diamètre du cercle dont leur Surface = Surface de l'objet
CONVEXMINANGLE Le minimum des angles formés par les paires de segments de ligne adjacentes qui constituent la limite polygonale d'un objet	**CROFTONPERIMETER** $\frac{1}{4}\iint n(\varphi.p)d\varphi dp$ Où $n(\varphi.p)$ est le nombre de pixels pour lesquels l'objet est intercepté par une ligne droite, avec un angle φ et une distance *p* du point d'origine.
CONVEXITY $\frac{\text{Surface de l'objet}}{\text{Surface de son enveloppe convexe}}$	**HEIGHT** Hauteur
HOLESTOTALAREA Surface des trous	**MBR.WIDTH** Largeur du MBR
LOGOFHEIGHTTOWIDTH $\text{Log}_{10}\left(\frac{\text{Hauteur}}{\text{Largeur}}\right)$	**NUMBEROFBLOBS** Nombre de composantes « 4-connecté » dans un objet
MAJORAXIS Direction de l'axe principal	**NUMBEROFHOLES** Nombre de trous
MBR.AREA Surface du MBR	**PERIMETER** Périmètre
MBR.FILLRATIO $\frac{\text{Surface}}{\text{MBR. Area}}$	**MBR.HEIGHT** Hauteur du MBR
PERIMETERVARIATION Somme des variations de direction entre les pixels de la limite	**SYMMETRYMEANDIFFERENCE** $\frac{1}{N}\sum_{i=0}^{N}\lvert p(i) - p'(i)\rvert$ Où N = Nombre de pixels de la limite/2 $p(i) = \left\|\overrightarrow{CX}\right\|, p'(i) = \left\|\overrightarrow{CX'}\right\|$ $Angle(\overrightarrow{CX}, \overrightarrow{CX'}) = \pi$
MBR.HEIGHTWIDTHRATIO $\frac{\text{MBR. Height}}{\text{MBR. Width}}$	**WIDTH** Largeur

De plus, sur les 13 097 objets de l'échantillon d'étude, l'appartenance de chaque objet à la catégorie « bâti » ou « non bâti » a été vérifiée visuellement. Autrement dit, il s'agissait d'étiqueter les objets (Nagao *et al.*, 1979b) dans le but d'avoir une « base de connaissances » (*Knowledge-based* en anglais) (Matsuyama et Hwang, 1990 ; Forestier *et al.*, 2012) sur les bâtiments, permettant ainsi de mettre au point l'analyse de ces données. Rappelons que notre objectif n'est pas de construire « un système à base de connaissances (SBC) » (Matsuyama et Hwang, 1990 ; Forestier *et al.*, 2012). Nous voulons approfondir les connaissances sur les paramètres morphologiques qui permettront potentiellement une différenciation « bâti/non bâti ». Certes, il est possible que ces paramètres morphologiques soient pertinents dans la construction d'un SBC, notamment dans les descripteurs de l'objet, mais tel n'est pas l'objectif de cette recherche. Nous explorons la « classification fondée sur l'objet » comme un choix parmi d'autres pour discriminer les bâtiments, dans le cadre de l'utilisation des données télédétection à THRS pour l'estimation de la population à des fins épidémiologiques.

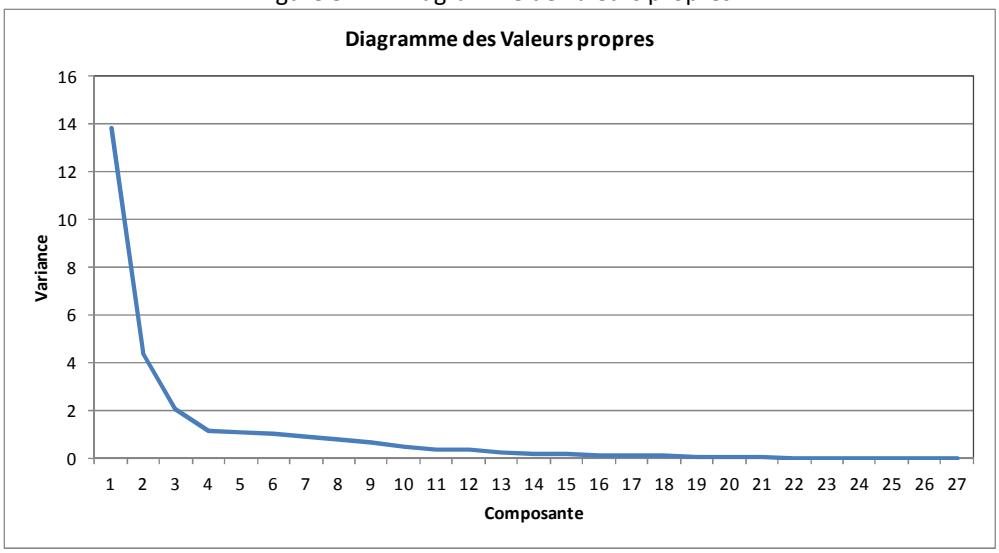
Pour l'analyse des données, nous nous sommes inspirée des travaux de Guerois (2003) ainsi que de ceux d'Ackermann et Mering (2007). Ces travaux se sont intéressés aux indicateurs morphologiques pour caractériser les villes, et ont conclu que la combinaison de différents paramètres de forme apporte plus de connaissances dans l'appréhension de la morphologie des villes que l'utilisation des paramètres pris séparément. Pour se faire, ils ont utilisé l'analyse en composantes principales (ACP). L'ACP permet d'affiner la description de données, d'éliminer les bruits (filtrer les données aberrantes), et de révéler des interactions et des associations entre les données (Sanders, 1989).

Sur le plan pratique, nous avons obtenu un tableau de données avec 13 097 « unités statistiques » (objets), et 30 « variables ». Parmi ces variables 28 correspondent aux attributs morphologiques (tab. 3-2) ; une représente le groupe « bâti/non bâti » ; et une autre identifie l'appartenance au niveau d'analyse « petit, moyen, grand »- « ouverture/fermeture » ; ces deux dernières variables sont utilisées comme éléments « supplémentaires[40] » dans l'analyse des données (fig. 3-22). L'analyse s'est effectuée à l'aide du logiciel R 2.13. avec le package ade4 (Chessel *et al.*, 2004). Pour ce faire, une ACP

[40] Les éléments supplémentaires sont des unités pour les quelles on dispose des observations mais dont on ne veut pas tenir compte dans le calcul des paramètres statistiques (Foucart, 1997)

a été appliquée au tableau de données [13 097, 28]. L'inertie expliquée pour les deux premières composantes atteint 64,66% de la variance totale des données (fig. 3-21). La contribution des individus sur ces deux premiers axes (fig. 3-22a) a permis de les filtrer, en éliminant les valeurs « extrêmes / aberrantes ». Ainsi nous avons retenu 11 761 individus (fig. 3-22b).

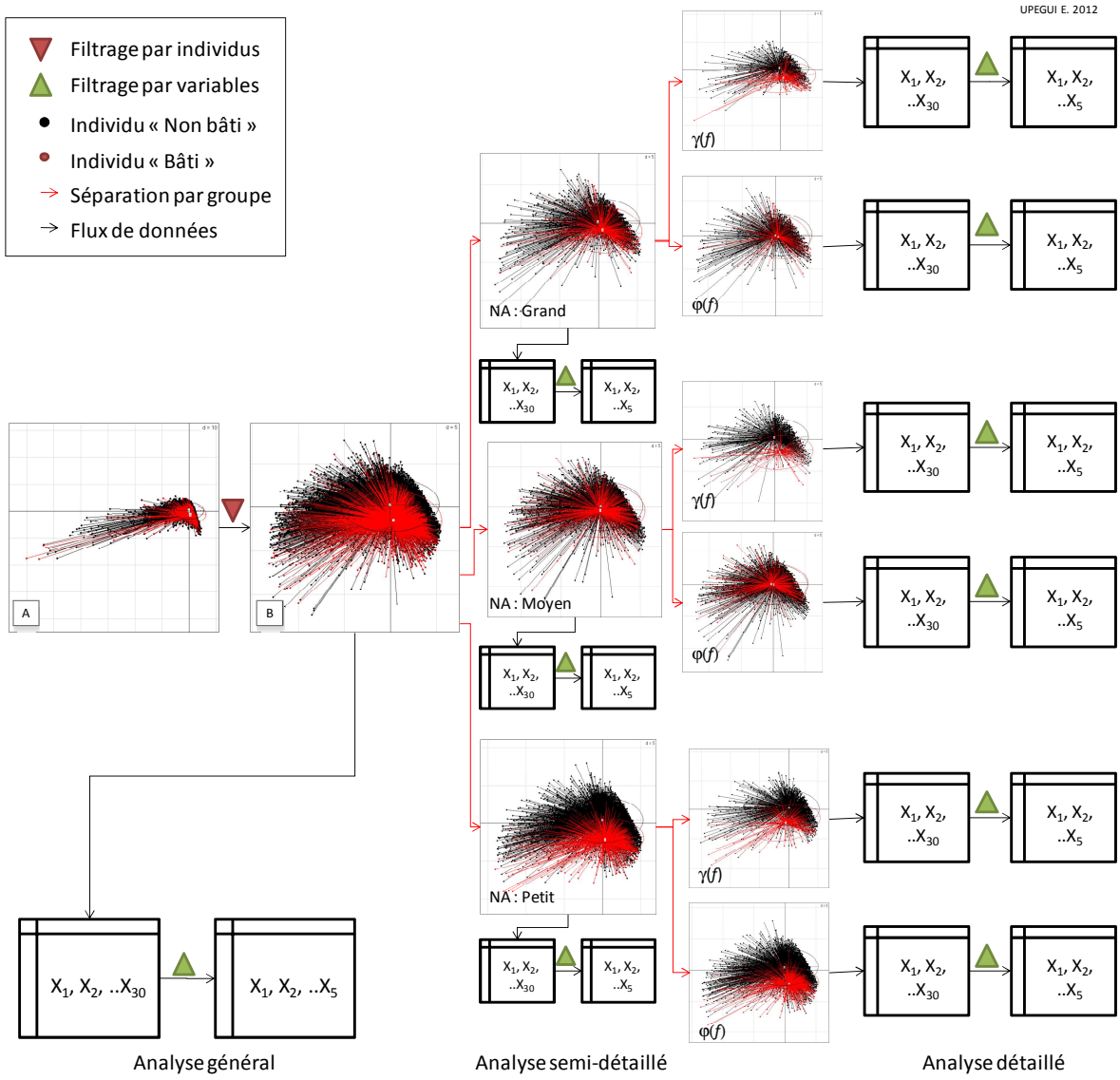
Figure 3-21 : Diagramme de valeurs propres



Ensuite l'analyse des données a été faite à trois niveaux (fig. 3-22) : le premier niveau -général- sur tous les individus, tous niveaux d'analyses confondus ; le deuxième niveau -semi-détaillé- sur les niveaux d'analyse « petit, moyen, grand », sans différenciation par teinte, avec trois jeux de données au total ; le dernier niveau -détaillé- sur les groupes d'individus classés par niveau d'analyse « petit, moyen, grand » et par teinte « ouverture/fermeture », avec un total de six jeux de données.

Postérieurement, une ACP a été appliquée à nouveau sur chacun des dix jeux de données, cette fois afin d'analyser la contribution des variables. Nous décrivons en détail la procédure pour le jeu de données « général » (11 761 individus), la même étant suivie pour les neuf jeux de données restants.

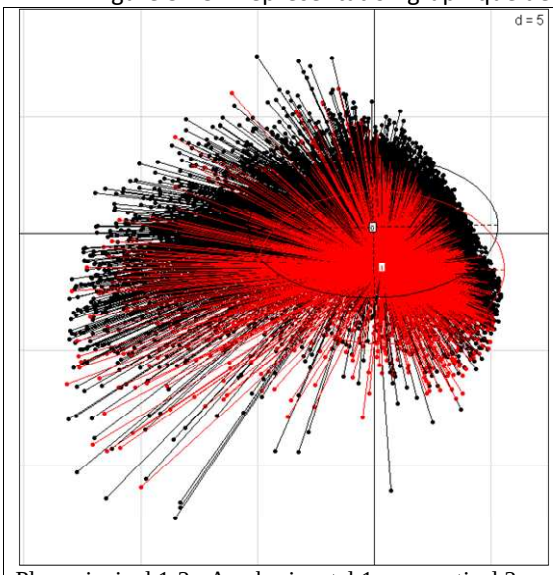
Figure 3-22 : Diagramme du flux de données pour l'ACP



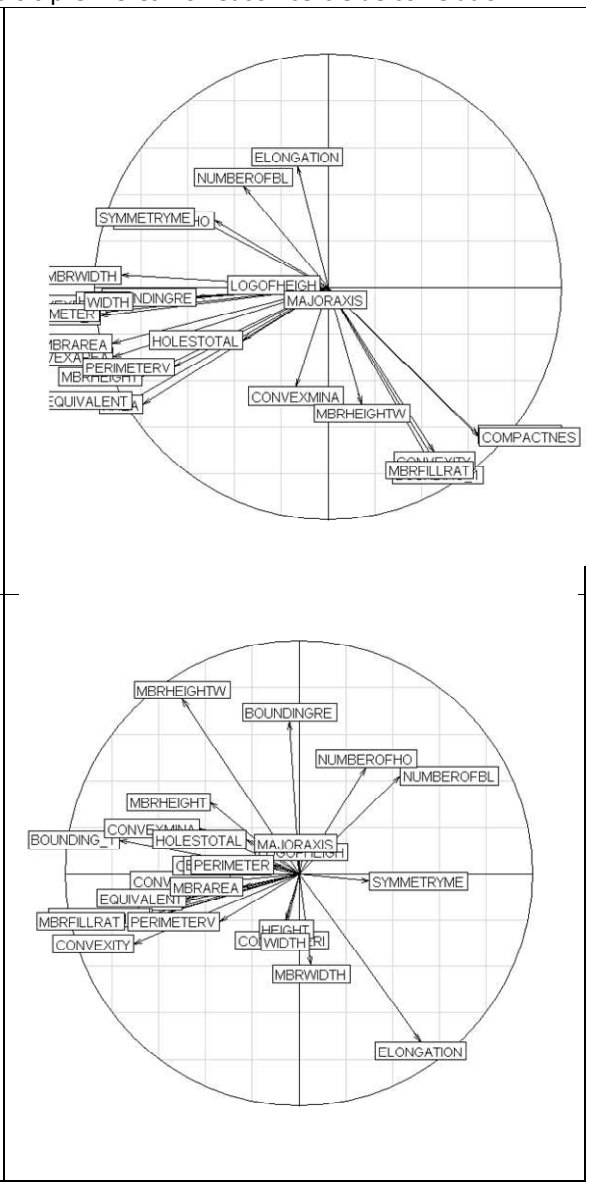
L'analyse a été centrée sur les 10 premiers axes, qui représentent 95,41% de la variance. Le pourcentage de variance cumulée indiqué sur ces axes est le suivant : ACP_1=45,37%, $ACP_{1:2}$=61,48%, $ACP_{1:3}$=71,09%, $ACP_{1:4}$=75,99%, $ACP_{1:5}$=80,68%, $ACP_{1:6}$=84,30%, $ACP_{1:7}$=87,86%, $ACP_{1:8}$=91,00%, $ACP_{1:9}$=93,53%, $ACP_{1:10}$=95,41%. La représentation graphique des trois premiers axes de l'ACP et leurs cercles de corrélation sont illustrés dans la figure 3-23 (la totalité des graphiques apparaissent dans l'annexe 3). En analysant ces graphiques, nous observons que l'axe 2 (16,11% de l'inertie expliquée) permet de séparer le centre de gravité des classes « bâti » (en rouge) / « non bâti » (en noir). Quant à l'axe 1, même s'il explique 45,37% de l'inertie, il ne discrimine pas « bâti/non bâti ». La variance expliquée par cet axe correspond peut-être à la diversité de classes qui

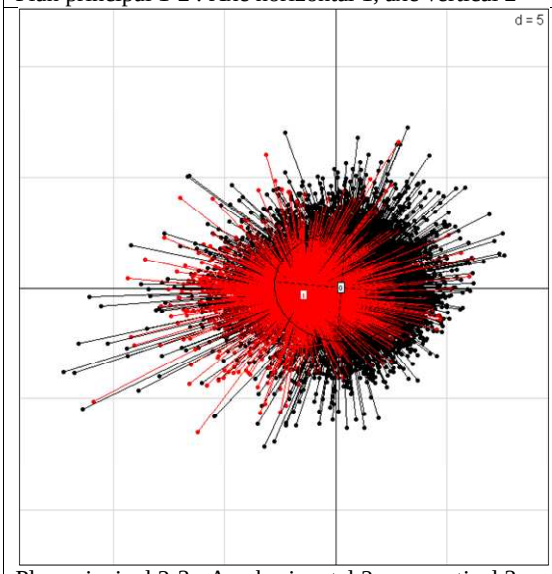
constituent la catégorie « non bâti » : par exemple les voies, les voitures, les cours intérieures, les trottoirs, les ombres.

Figure 3-23 : Représentation graphique des trois premières ACP et son cercle de corrélation



Plan principal 1-2 : Axe horizontal 1, axe vertical 2





Plan principal 2-3 : Axe horizontal 2, axe vertical 3

L'étape suivante vise l'analyse de la contribution des 28 variables sur chacun des axes, notamment l'axe 2 qui s'est révélé comme un discriminateur de « bâti/non bâti ». Le tableau 3-3 contient ces contributions qui ont été organisées de manière décroissante par rapport à l'axe 2. Nous avons ainsi mis en évidence les variables qui « expliquent » la majorité de l'inertie dans l'axe ; et nous avons retenu les cinq premières. Ces variables (soulignées en gris sur le tableau) sont : BOUNDINGRECTFILL (surnommé *BOUNDING_1* dans le tableau) ; MBR.FILLRATIO (*MBR.FILLR*) ; CONVEXITY (*CONVEXITY*) ; COMPACTNESS (*COMPACTNES*) ; et CIRCULARITY (*CIRCULARIT*).

Tableau 3-3 : Contribution relative des variables sur chacun des 10 axes de l'ACP

	Contribution relative (par 10 000)										
Variable	Axe1	Axe2	Axe3	Axe4	Axe5	Axe6	Axe7	Axe8	Axe9	Axe10	con.tra
BOUNDING_1	2217	-5905	215	-90	-179	7	-16	105	25	371	357
MBRFILLRAT	1891	-5579	-421	-432	-677	197	0	-48	1	259	357
CONVEXITY	2072	-4996	-922	-432	-652	107	-4	1	22	172	357
COMPACTNES	4185	-4105	-359	-146	-163	-344	1	31	1	-240	357
CIRCULARIT	4141	-3971	-358	-174	-230	-356	4	3	1	-352	357
AREA	-6276	-2489	-143	-16	-13	-2	-1	9	7	-2	357
MBRHEIGHTW	208	-2489	5624	163	373	-534	16	64	0	-62	357
EQUIVALENT	-7042	-2328	-108	-15	-15	-6	0	0	1	20	357
CONVEXMINA	-186	-1806	383	60	6	64	202	-4912	-2333	-23	357
MBRHEIGHT	-6364	-1446	947	82	170	-307	2	37	2	-46	357
PERIMETERV	-4312	-1154	-424	564	357	407	0	-514	1067	-122	357
CONVEXAREA	-8453	-873	-8	4	18	-101	0	18	2	-40	357
MBRAREA	-8545	-570	-30	12	34	-201	0	44	4	-71	357
HOLESTOTAL	-1323	-503	223	-11	-64	3928	-54	1683	-1114	-1064	357
CROFTONPER	-9492	-144	13	-2	-5	41	-1	5	2	65	357
EQUIVALE_1	-9492	-144	13	-2	-5	41	-1	5	2	65	357
PERIMETER	-9567	-115	15	-5	-19	34	0	0	5	39	357
CONVEXPERI	-8981	-52	-616	-2	0	-89	0	1	0	-1	357
WIDTH	-6978	-34	-649	1031	-640	-151	3	0	-8	-7	357
HEIGHT	-7156	-30	-405	-1341	550	-21	0	-24	3	-20	357
BOUNDINGRE	-3176	-17	4273	5	-15	930	-5	0	57	1252	357
MAJORAXIS	-1	-2	86	1	-5	-69	-9622	-189	-17	-7	357
LOGOFHEIGH	-10	1	32	-6322	3290	61	-1	-51	24	-9	357
MBRWIDTH	-7857	27	-1555	-17	-18	-36	-1	0	0	0	357
NUMBEROFHO	-2367	824	2069	-1035	-2728	53	10	-149	225	-83	357
SYMMETRYME	-3271	913	-9	-701	-189	-1451	1	527	-1907	408	357
NUMBEROFBL	-1308	1882	1771	-995	-2513	-165	30	-233	242	-455	357
ELONGATION	-172	2713	-5237	-67	-193	433	-10	-135	-11	9	357

Le tableau 3-4 contient les cinq variables dont la contribution est la plus importante dans la discrimination « bâti/non bâti », dans chacun des 10 jeux de données analysés. La variable *BOUNDING_1* apparait comme déterminante dans tous les niveaux d'analyses, c'est-à-dire qu'elle est indépendante de la taille et de la teinte de l'objet. Quant au niveau d'analyse « petit », les variables discriminantes sont les mêmes sans tenir compte de la teinte, à savoir : *MBRFILLRAT, CONVEXITY, BOUNDING_1, COMPACTNES,* et *CIRCULARIT.* Pour le niveau d'analyse « grand » les variables *ELONGATION* et *MBRHEIGHTW* (*MBR.HEIGHTWIDTHRATIO)* apparaissent comme discriminantes, permettant peut-être de séparer les bâtiments des routes. Les autres variables discriminantes pour les objets

« grands », au niveau semi-détaillé sont : *MBRFILLRAT*, et *CONVEXITY*. Quant à la différenciation pour le critère « teinte » : pour les objets foncés (fermeture) la *MBRHEIGHT*, et la *COMPACTNES* sont discriminantes ; tandis que pour les objets clairs (ouverture) ce sont la *MBRFILLRAT*, et la *CONVEXITY*. Pour le niveau d'analyse « moyen », la variable *ELONGATION* différencie les objets foncés tandis que la *CIRCULARIT* différencie les clairs. Le reste des variables est commun aux deux niveaux, à savoir : *BOUNDING_1, MBRFILLRAT, CONVEXITY*, et *COMPACTNES.*

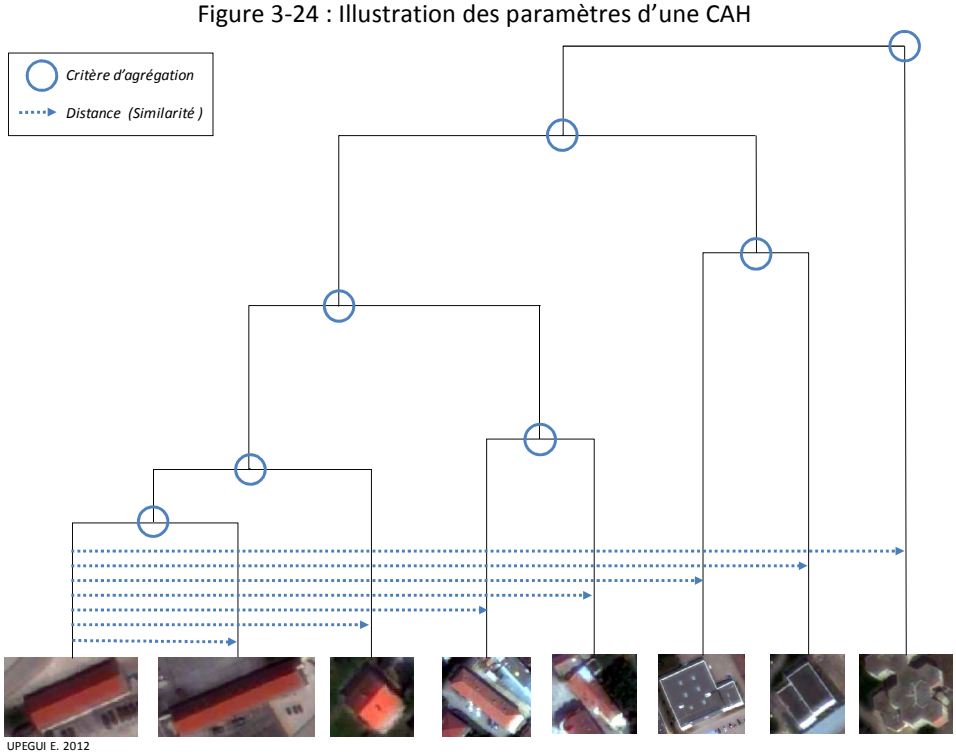
Tableau 3-4 : Variables discriminantes pour l'identification bâti/non bâti, dans les différents niveaux d'analyse

	Variable 1	**Variable 2**	**Variable 3**	**Variable 4**	**Variable 5**
Niveau Général	BOUNDING_1	MBRFILLRAT	CONVEXITY	COMPACTNES	CIRCULARIT
- N.A Grand	BOUNDING_1	MBRHEIGHTW	ELONGATION	MBRFILLRAT	CONVEXITY
✓ **N.A Grand Fermeture**	BOUNDING_1	MBRHEIGHTW	ELONGATION	MBRHEIGHT	COMPACTNES
✓ **N.A Grand Ouverture**	BOUNDING_1	MBRHEIGHTW	ELONGATION	MBRFILLRAT	CONVEXITY
- N. A Moyen	BOUNDING_1	MBRFILLRAT	CONVEXITY	COMPACTNES	CIRCULARIT
✓ **N. A Moyen Fermeture**	BOUNDING_1	MBRFILLRAT	ELONGATION	CONVEXITY	COMPACTNES
✓ **N. A Moyen Ouverture**	BOUNDING_1	MBRFILLRAT	CONVEXITY	COMPACTNES	CIRCULARIT
- N.A. Petit	MBRFILLRAT	CONVEXITY	BOUNDING_1	COMPACTNES	CIRCULARIT
✓ **N.A. Petit Fermeture**	MBRFILLRAT	CONVEXITY	BOUNDING_1	COMPACTNES	CIRCULARIT
✓ **N.A. Petit Ouverture**	MBRFILLRAT	CONVEXITY	BOUNDING_1	COMPACTNES	CIRCULARIT

3.2.3 Classification ascendante hiérarchique : une méthode non dirigée

La classification ascendante hiérarchique (CAH) est une méthode de classification non dirigée, couramment utilisée dans l'analyse de données (Benzécri *et al.*, 1980). Cette méthode est intégrée fréquemment à l'ACP afin de réaliser des typologies sur les objets analysés. C'est pour cette raison que nous l'avons choisie pour poursuivre la démarche de classification fondée sur l'objet. D'ailleurs, c'est la méthode employée par Ackermann et Mering (2007), dont nous nous sommes inspirée. A la différence d'autres méthodes de classification non dirigée (par exemple ISODATA ou K-mens, méthodes de répartition), dans la CAH le nombre de classes désiré ne se fixe pas. La CAH se fonde sur l'agrégation progressive des objets considérés (Sanders, 1989). Au début du processus, chaque objet constitue une classe en soi ; puis les objets se regroupent progressivement deux à deux, à l'aide d'un critère d'agrégation (distance entre classes), en fonction de leur similarité, jusqu'à constituer une seule classe contenant tous les objets. La similarité se calcule en fonction d'une distance mesurée entre les objets (matrice des distances), permettant de construire l'agrégation hiérarchique des objets pour former les classes. Cette agrégation peut se représenter graphiquement par un « dendrogramme » ou diagramme en arbres (fig. 3-24).

Figure 3-24 : Illustration des paramètres d'une CAH



L'algorithme de classification de la CAH se déroule en trois étapes (Chessel *et al.*, 2004b) :

1. Utiliser la plus petite distance (consignée dans la matrice des distances calculées entre les objets) pour former un couple. Cela se fait sur la totalité des n objets, n-1 objets restant ainsi.

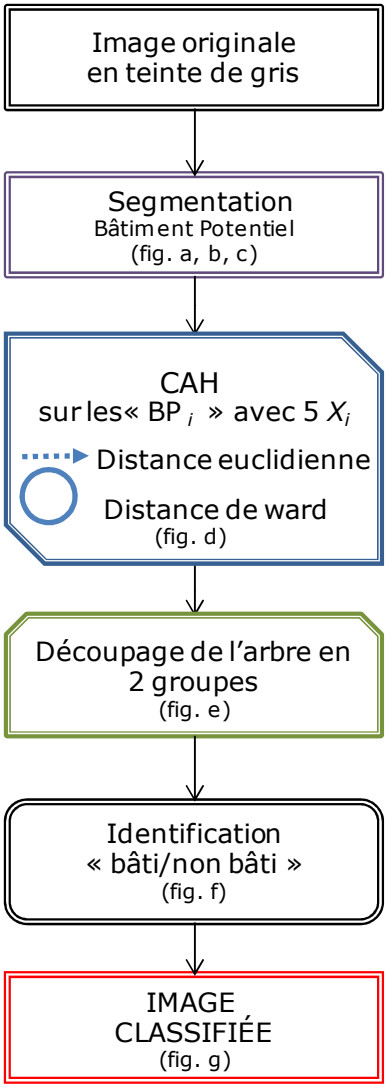
2. Regrouper les valeurs h(i), associées à chaque objet, selon le critère M, puis assigner au nouveau groupe, une nouvelle valeur h supérieure à celles des deux objets regroupés.

3. Répéter les étapes 1 et 2 jusqu'à ce qu'il ne reste plus que la classe regroupant tous les objets avec une valeur supérieure à toutes les autres.

Néanmoins, chaque procédé qui définit M et h, donne une CAH particulière. Ainsi pour n objets, le nombre de hiérarchies possibles est : $\frac{(2n-3)!}{2^{n-2}\,(n-2)!}$

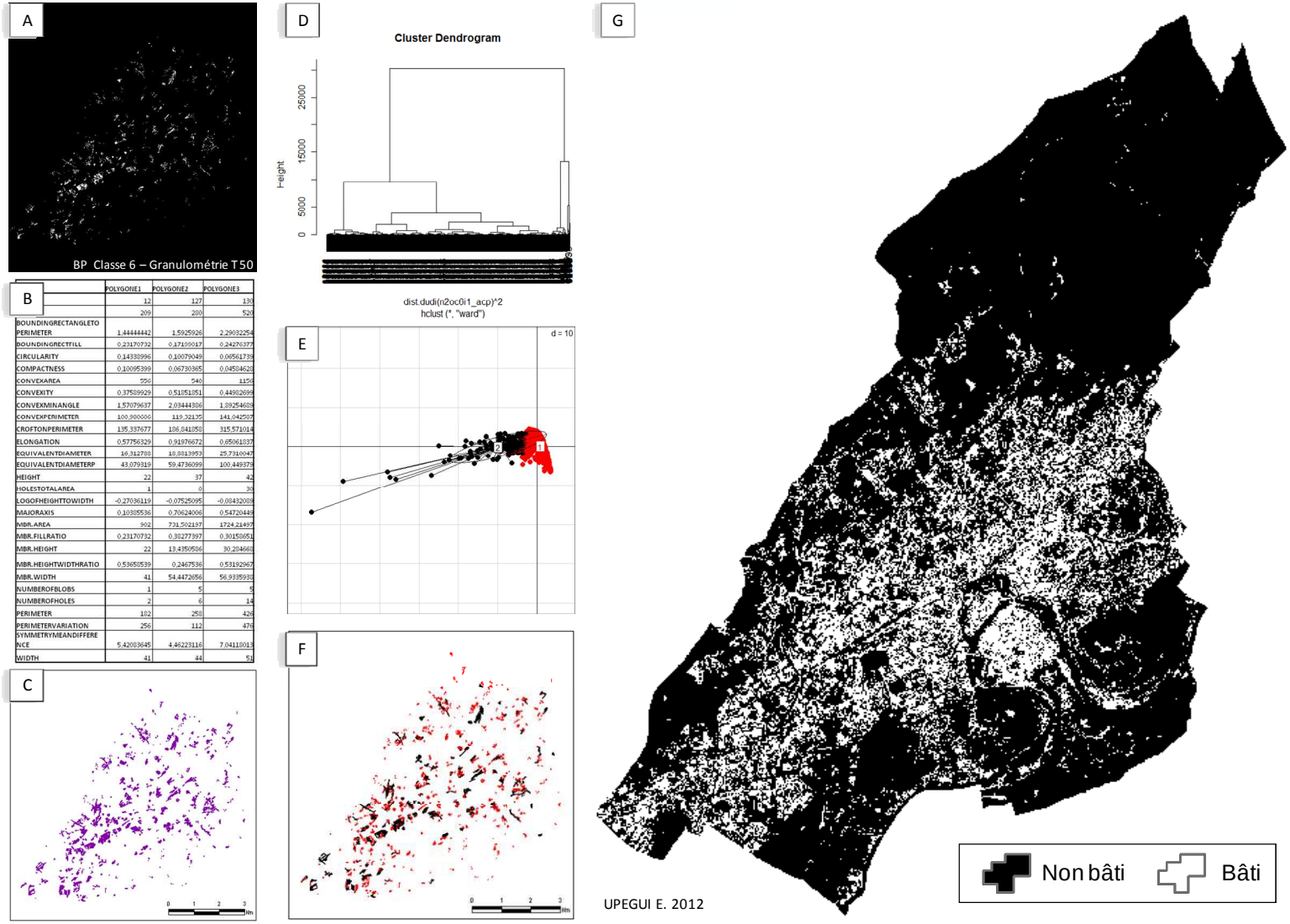
Abordons l'application de cette classification à notre zone d'étude (fig. 3-25) : on peut dire que la première étape de la classification -c'est-à-dire l'extraction des caractéristiques- a correspondu à la segmentation et l'analyse préliminaire des variables. Puis, la CAH a été appliquée sur chacune des 74 couches de bâtiments potentiels (BP) identifiées lors de la segmentation (cf. 3.2.2), en utilisant les 5 variables (X_i) pertinentes relevées dans le niveau d'analyse détaillé (cf. 3.2.3 : tab. 3-3). Comme critères, nous avons utilisé, d'une part, la distance euclidienne, pour la distance entre individus ; et d'autre part, la distance de Ward qui vise à maximaliser l'inertie interclasse (et donc à minimiser l'inertie intra-classe), pour la distance entre classes. Ensuite chacun des arbres issus de la CAH a été « coupé » en deux pour avoir deux groupes à étiqueter (troisième étape de la classification), visuellement, comme « bâti » ou « non bâti ».

Figure 3-25 : Algorithme de la classification ascendante hiérarchique



Du point de vue pratique, les bâtiments potentiels ont été créés par ENVI, par le biais d'une image en teinte de gris (*.tiff) (fig. 3-26a). Sur cette image deux procédures ont été suivies. D'un côté, sur Aphelion, les objets ont été créés (fig. 3-26b), et ils ont été exportés en format xls. De l'autre, sur ArcGIS, l'image a été vectorisée (fig. 3-26c) tout en gardant l'identifiant créé par Aphelion. Puis sur les tableaux de données des objets créés, la CAH a été effectuée (fig. 3-26d) à l'aide du logiciel R 2.13.1. Postérieurement chacun des arbres issus de la CAH a été « coupé » en deux groupes (fig. 3-26e). Ensuite, les polygones des objets ont été joints à la CAH à travers ses identifiants (fig. 3-26f) en utilisant ArcGIS. Enfin, une mosaïque a été réalisée avec tous les objets appartenant aux classes « bâties » (fig. 3-26g).

Figure 3-26 : Illustration de différentes étapes de l'algorithme de la classification ascendante hiérarchique pour l'extraction des bâtiments

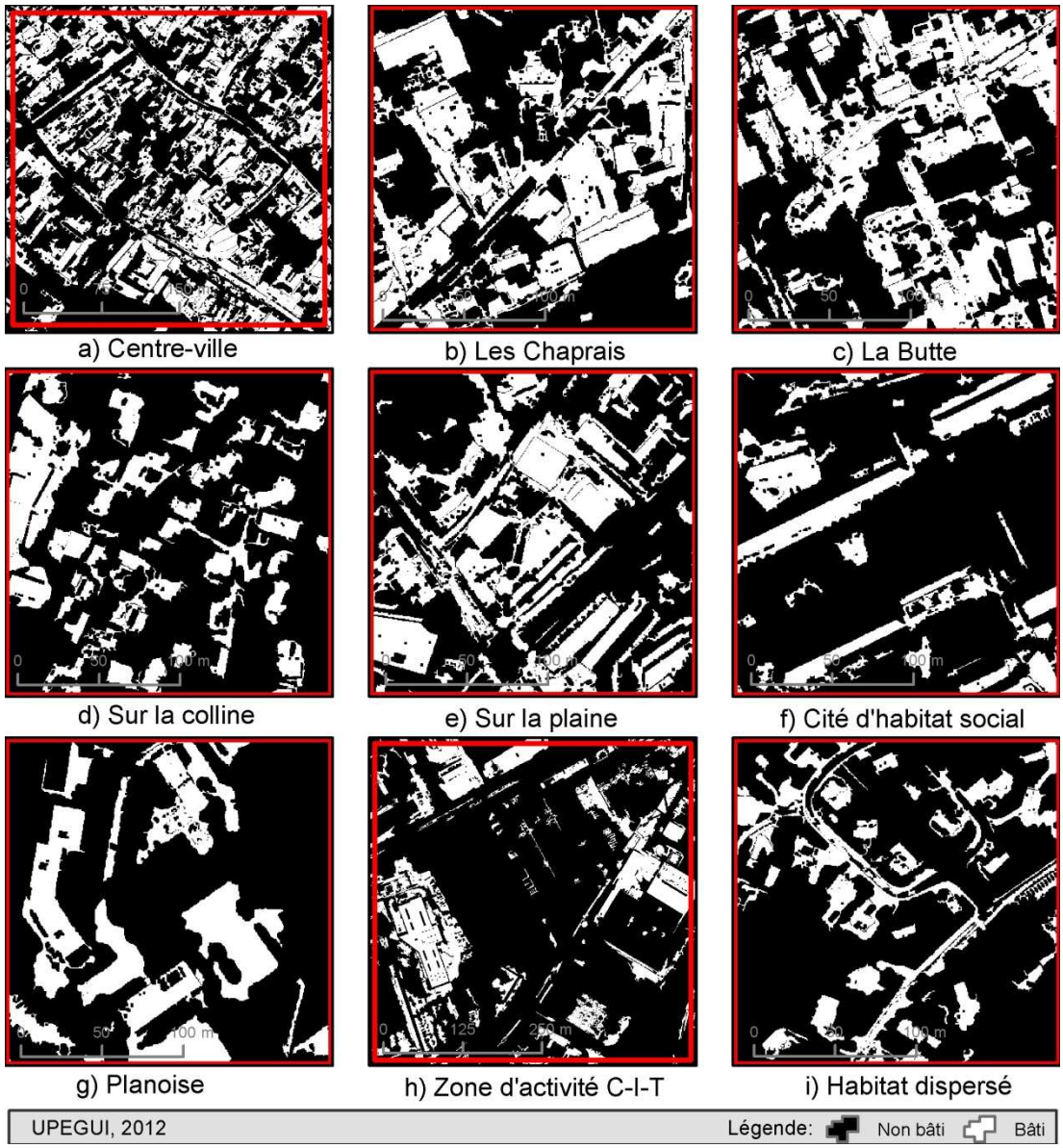


	POLYGONE1	POLYGONE2	POLYGONE3
	12	127	130
	209	280	520
BOUNDINGRECTANGLETOPERIMETER	1,44444442	1,5925926	2,29032254
BOUNDINGRECTFILL	0,23170732	0,17199017	0,24276377
CIRCULARITY	0,14338996	0,10079049	0,06561739
COMPACTNESS	0,10095399	0,06730365	0,04584628
CONVEXAREA	556	540	1150
CONVEXITY	0,37589929	0,51051051	0,44982699
CONVEXMINANGLE	1,57079637	2,03444386	1,89254689
CONVEXPERIMETER	100,900006	119,32135	141,042587
CROFTONPERIMETER	135,337677	186,841058	315,571014
ELONGATION	0,57756329	0,91976672	0,65061837
EQUIVALENTDIAMETER	16,312788	18,8813953	25,7310047
EQUIVALENTDIAMETERP	43,079319	59,4736099	100,449379
HEIGHT	22	37	42
HOLESTOTALAREA	1	0	30
LOGOFHEIGHTTOWIDTH	-0,27036119	-0,07525095	-0,08432089
MAJORAXIS	0,10305536	0,70624006	0,54720449
MBR.AREA	902	731,502197	1724,21497
MBR.FILLRATIO	0,23170732	0,38277397	0,30150651
MBR.HEIGHT	22	13,4350586	30,204668
MBR.HEIGHTWIDTHRATIO	0,53658539	0,2467536	0,53192967
MBR.WIDTH	41	54,4472656	56,9335938
NUMBEROFBLOBS	1	5	5
NUMBEROFHOLES	2	6	14
PERIMETER	102	258	426
PERIMETERVARIATION	256	112	476
SYMMETRYMEANDIFFERENCE	5,42003645	4,46223116	7,04110013
WIDTH	41	44	51

Analysant en détail les résultats de la classification ascendante hiérarchique dans chacune des typologies des quartiers (fig. 3-27), nous remarquons que dans le centre-ville (fig. 3-27a) l'extraction n'a pas été faite par pâtés de maisons. En effet, différentes toitures (ou pans des toitures) à taille et à formes irrégulières apparaissent dans la classification. Cependant, toutes les toitures n'ont pas été extraites, et certains tronçons de routes ont été classés comme « bâti ». Ce « comportement » se maintient globalement pour toutes les typologies. Aux Chaprais (fig. 3-27b), la plupart des toitures ont été clairement extraites. Dans les cités d'habitat social (fig. 3-27f), sur la plaine (fig. 3-27e), et à Planoise (fig. 3-27g) au moins un des pans des toitures a été extrait, en particulier celui qui est ensoleillé, et la plupart des routes ont été classées dans la catégorie « non bâti ». En revanche, certaines ombres des immeubles n'ont pas été complètement détachées de l'immeuble. Quant à l'habitat dispersé (fig. 3-27i), une grande partie des toitures a été bien extraite. Toutefois, la voie routière qui connecte ces maisons individuelles a été également classée dans la catégorie « bâti ». Cela n'est pas le cas sur la colline (fig. 3-27d) où la voie routière a été bien classée dans « non bâti ». À La Butte (fig. 3-27c), quasiment toutes les routes, reliant les bâtiments, ont été classées dans la catégorie « bâti ». Enfin, dans la zone d'activité C-I-T (fig. 3-27h), presqu'aucun bâtiment n'a été correctement extrait, tandis que certains parkings et quasi toutes les voies ont été bien classés dans « non bâti ». A noter que quelques voitures, sur les parkings, apparaissent dans les éléments classés comme « bâti ».

En résumé, la segmentation des objets par morphologie mathématique liée à la caractérisation des attributs morphologiques, notamment ceux révélés comme discriminants, permettent l'identification des bâtiments. Même si tous les bâtiments ne sont pas extraits (à l'inverse de la classification ISODATA, où on est dans une configuration spatiale de surestimation), la séparation des « bâtiments » des autres zones imperméables est assez bonne dans l'ensemble de la zone d'étude.

Figure 3-27 : Détails des échantillons classifiés par la classification ascendante hiérarchique



3.2.4 Régression logistique : une méthode dirigée

Le modèle de régression logistique (RL) (ou modèle logistique) est un modèle multi-varié qui permet d'exprimer sous forme de probabilité (π_i) la relation entre une variable Y dichotomique (deux valeurs possibles 0 ou 1) et une ou plusieurs variables *Xi*, qui peuvent être qualitatives ou quantitatives (Berkson, 1944; Hastie *et al.*, 2008). Ainsi, la probabilité π_i est définie par :

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_i+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_1 x_i+\ldots+\beta_k x_k}}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_k$$

L'évaluation de ce modèle peut se faire avec différents critères, notamment le critère d'AIC (*Akaike Information Criterion,* en anglais - Akaike, 1974) :

$$AIC = -2\log(L) + 2k$$

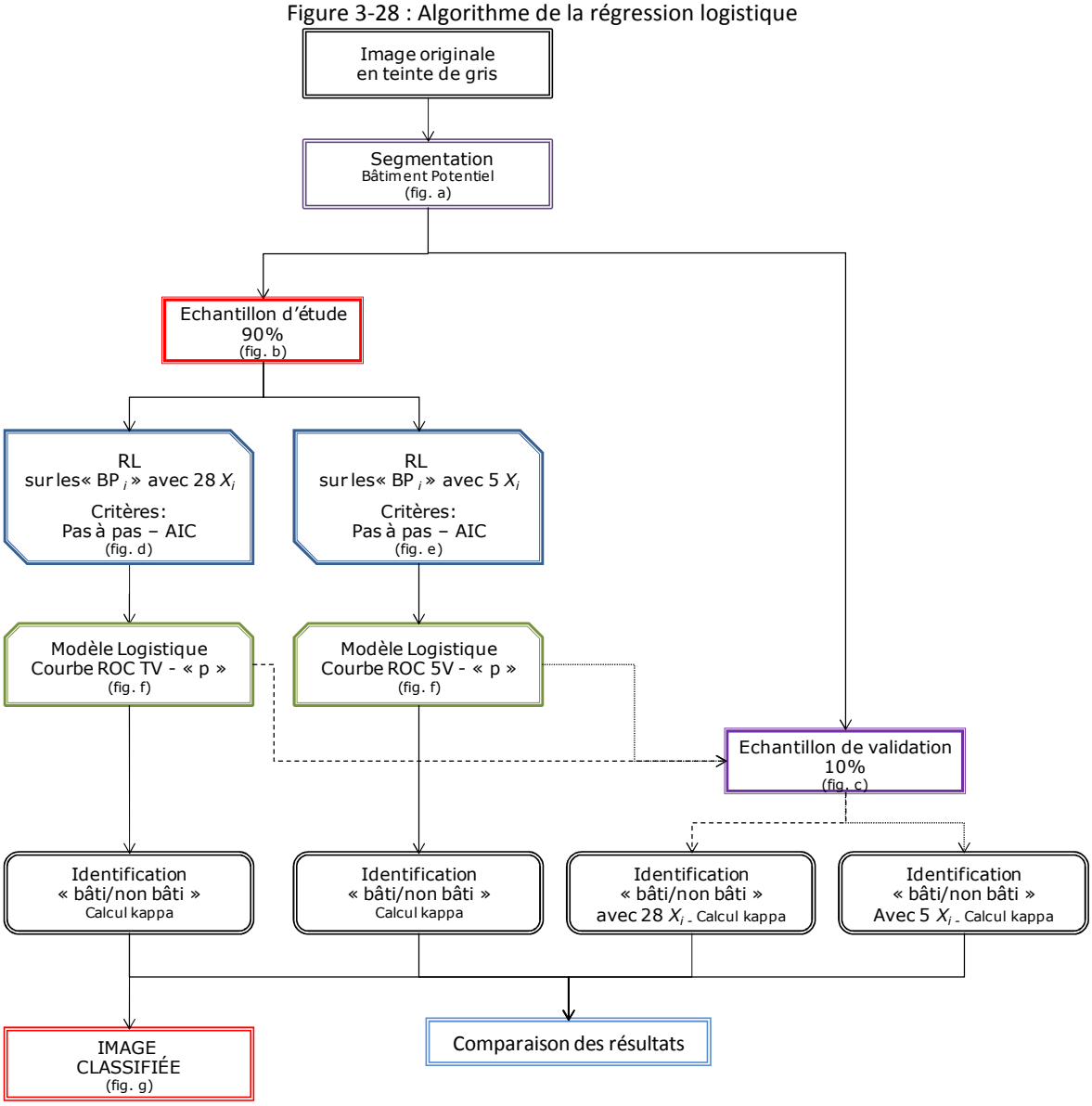
Où L est la vraisemblance maximisée et k le nombre de paramètres (variables) dans le modèle. Le meilleur modèle correspond à la valeur la plus basse d'AIC.

Etant donné nos jeux de données -une variable dichotomique, « bâti » / « non bâti » ; et un groupe de variables quantitatives, attributs morphologiques- la RL se présente comme une alternative « déterministe » peu dépendante de l'opérateur (une fois l'étape d'apprentissage achevée) pour l'extraction des bâtiments dans le cadre d'une stratégie fondée sur l'objet. Le caractère « déterministe » de ce modèle de régression s'ajuste aux contraintes (simplicité, automatisme, reproductibilité, et exportabilité) fixées pour l'estimation de la population dans une optique de santé publique (prévision et action sanitaire) ; et d'ailleurs la RL est un outil majeur en épidémiologie. C'est pour ces raisons que nous avons choisi cette méthode dirigée pour extraire les bâtiments.

Dans notre cas, le but de la RL est de déterminer la probabilité π_i pour un objet d'être « bâti » à partir des attributs morphologiques, et le seuil avec lequel la plupart des objets « bâtis » peuvent être extraits. La performance de la RL dans la « prédiction » des bâtiments à été testée selon un protocole de vérification (fig. 3-28). L'échantillon d'étude (10% de bâtiments potentiels - cf. 3.2.1) a été divisé en deux : 90% sur lequel les modèles

sont construits (échantillon d'étude), et les 10% restants sur lequel le modèle est testé (échantillon de validation). En outre, afin de vérifier la performance des variables pertinentes pour la discrimination « bâti » / « non bâti » identifiées auparavant (cf. 3.2.2), nous avons décidé d'appliquer la RL, en parallèle, à deux groupes de variables : d'une part, tous les attributs morphologiques (TV, à savoir les 28 attributs) ; et d'autre part, les 5 variables pertinentes (5V) identifiées par chaque niveau d'analyse.

Figure 3-28 : Algorithme de la régression logistique



Concrètement, la régression s'est effectuée de manière automatique par R.2.13.1, par la « méthode pas à pas descendante » (*stepwise,* en anglais) en utilisant le critère d'AIC pour retirer les variables non significatives statistiquement du modèle. Néanmoins, dans certains cas, quelques variables ont été retirées du modèle de manière manuelle,

pour assurer cette significativité statistique. Deux régressions logistiques -une avec TV (fig. 3-30d) et une autre avec les 5V (fig. 3-30e)- ont été effectuées sur chacun des échantillons d'étude des 72 couches de BP issues de l'image GeoEye. Quant aux 2 couches de BP issues de l'ortho-photo, un modèle généralisé construit à partir des BP du niveau d'analyse « petit » a été appliquée. Une fois les modèles logistiques obtenus, la courbe ROC (Sing *et al*., 2005) (fig. 3-30f) a permis d'établir le seuil où π_i représente un bon compromis pour extraire les bâtiments (fig. 3-30g). Puis chaque modèle de régression logistique, calculé à l'aide de l'échantillon d'étude extrait de chacune des couches des BP donnée, a été appliqué sur la totalité des objets constituant ces couches. Puis la valeur du seuil a été appliquée aux objets, permettant ainsi l'identification « bâti » / « non bâti ». Enfin une mosaïque a été réalisée avec les classes « bâti » de toutes les couches.

En ce qui concerne la comparaison des résultats obtenus avec les différents modèles de régression logistique (fig. 3-29), une valeur « moyenne » du coefficient kappa[41] a été calculée avec les différentes valeurs de kappa obtenues pour chacun des modèles de régression dans chacun des niveaux d'analyse. Il faut signaler que dans certaines couches de l'échantillon de validation (10% de l'échantillon) il n'a pas été possible d'estimer la valeur de kappa car il n'y avait aucun objet de type « bâti », en raison du tirage aléatoire utilisé pour les échantillons. Au-delà, on observe que les valeurs les plus élevées du coefficient kappa (avec un accord modéré) ont été obtenues par les modèles de régression logistique calculés sur l'échantillon d'étude (90% d'objets) où toutes les variables interviennent. Toutefois, l'application de ces modèles sur l'échantillon de validation (10% des objets) n'a pas toujours eu le même comportement : les valeurs de kappa étant moindres que celles de l'échantillon d'étude et leur écart-type étant plus important. D'autre part, les modèles de régression logistique calculés sur l'échantillon d'étude (90% des objets) où seulement les cinq variables pertinentes interviennent, obtiennent des valeurs de kappa plus faibles mais avec des écarts-types également plus faibles. En revanche, ces valeurs sont plus stables dans l'application de ces modèles sur l'échantillon de validation. En outre, les valeurs les plus basses de kappa correspondent au niveau d'analyse « grand », plus particulièrement à l'ouverture (avec les écarts-types

[41]Indice statistique qui reflète la proportion de l'accord entre les données de référence et les données classifiées après avoir retiré l'accord dû au hasard (Cohen, 1960).

plus élevés), tandis que les valeurs les plus hautes correspondent au niveau d'analyse « petit » (avec des écarts-types plus petits). Les écarts plus importants dans les valeurs estimées par les différents modèles se trouvent dans les niveaux d'analyse « grand » et « moyen », notamment dans l'ouverture, tandis que les écarts plus petits correspondent au niveau d'analyse « petit » (ouverture et fermeture) et au niveau d'analyse « moyen », notamment dans la fermeture. On peut donc conclure que l'utilisation de toutes les variables dans le modèle permet l'identification des « particularités » dans les objets, bien que sa reproductibilité soit moindre. En revanche, l'utilisation des variables pertinentes dans la caractérisation des objets de différentes tailles, permet l'identification des bâtiments « plus standard », c'est-à-dire sans particularités (irrégularités de forme, taille, orientation, …) ; ce qui est plus reproductible. Ainsi, on constate que les variables identifiées, lors de l'étude préliminaire de la potentialité des indicateurs morphologiques pour la discrimination « bâti » / « non bâti », sont pertinentes pour l'identification des bâtiments, bien que cela n'explique pas toute la variabilité des bâtiments. Enfin, le modèle de RL qui tient compte de toutes les variables a été utilisé pour extraire les bâtiments. En effet, on est dans une optique de modélisation « explicative » pour identifier « bâti » / « non bâti », et donc on dispose de toutes les variables déjà mesurées.

Figure 3-29 : Comparaison des résultats obtenus avec les différents modèles de régression logistique. À gauche valeurs moyennes de kappa ; à droite valeurs écart-type moyennes

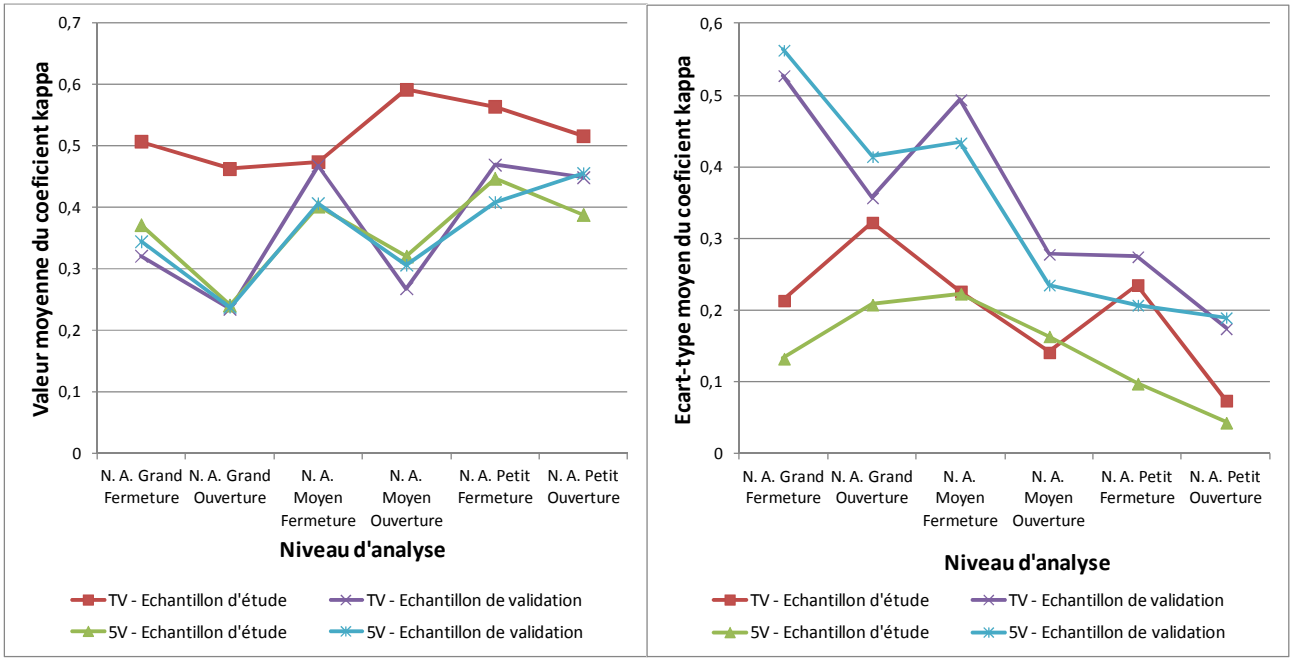
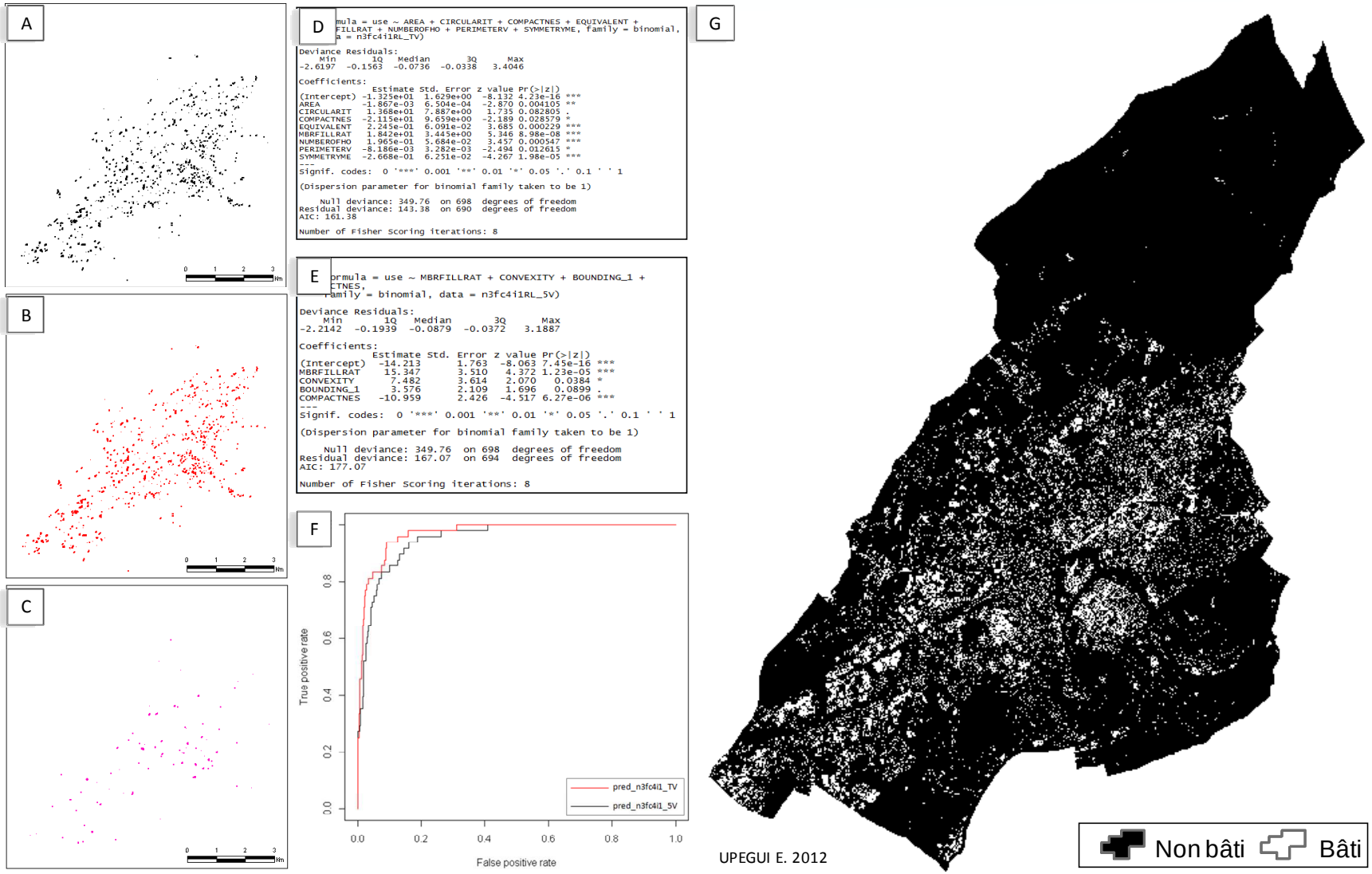


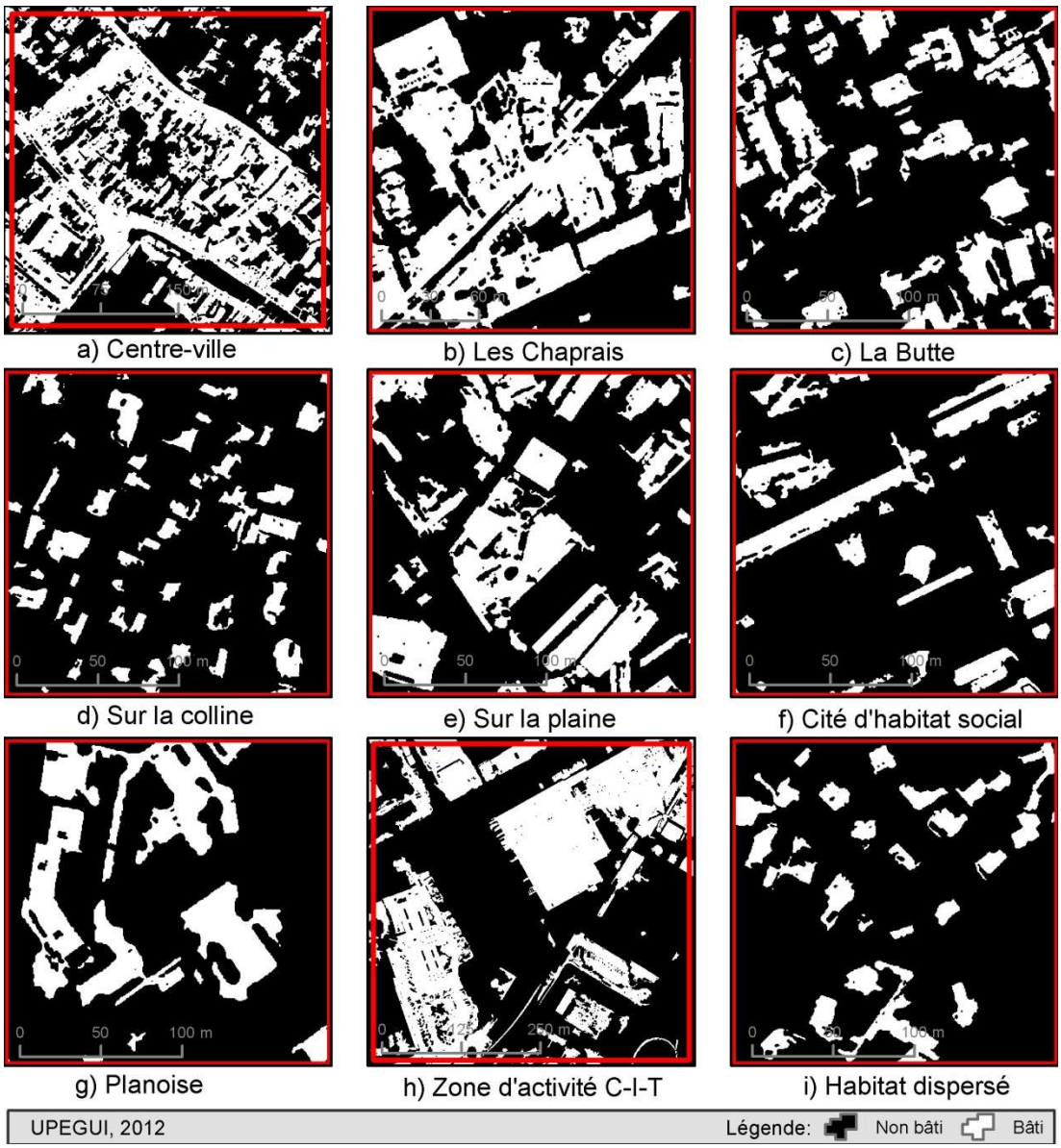
Figure 3-30 : Illustration de différentes étapes de l'algorithme de la régression logistique pour l'extraction des bâtiments



Après analyse détaillée des résultats de l'extraction des bâtiments par la régression logistique dans chacune des typologies des quartiers (fig. 3-31), nous remarquons que dans le centre-ville (fig. 3-31a) quelques rues relient les pâtés de maisons. Toutefois, toutes les toitures (ou pans des toitures) n'ont pas été extraites, en particulier celles non ensoleillées. Aux Chaprais (fig. 3-31b), la plupart des toitures ont été correctement extraites même si certaines routes relient les maisons. Dans les cités d'habitat social (fig. 3-31f), et sur la plaine (fig. 3-31e), au moins un de pans des toitures a été extrait, notamment celui qui est ensoleillé. Toutefois, certaines ombres projetées par des immeubles n'ont pas été complètement détachées de ces derniers. En revanche, les immeubles sont bien détachés des routes. À Planoise (fig. 3-31g) la plupart des immeubles ont été bien extraits, bien que quelques tronçons de routes soient classés comme « bâti ». Quant à l'habitat dispersé (fig. 3-31i), et sur la colline (fig. 3-31d) la plupart des toitures (ou pans des toitures) et des voies routières ont été bien classées comme respectivement « bâti » et « non bâti ». À La Butte (fig. 3-31c), une grande partie des toitures a été bien extraites, mais quelques zones imperméables ont été classées comme « bâti ». Enfin, dans la zone d'activité C-I-T (fig. 3-31h), quasiment tous les bâtiments ont été extraits ; et les routes ont été bien classées comme « non bâti ». En revanche, les parkings non pas pu être détachés des immeubles.

En bilan, la segmentation des objets par morphologie mathématique liée à l'extraction des bâtiments par régression logistique permet d'identifier les bâtiments. Bien que l'on soit dans un environnement de sous-estimation des surfaces bâties, car tous les bâtiments n'ont pas été extraits, on a une extraction relativement précise des « bâtis ». Cependant, l'identification des « bâtiments » à formes irrégulières, des « bâtiments » en habitat dispersé, et la séparation des « bâtiments » des routes, est assez bonne dans l'ensemble de la zone d'étude.

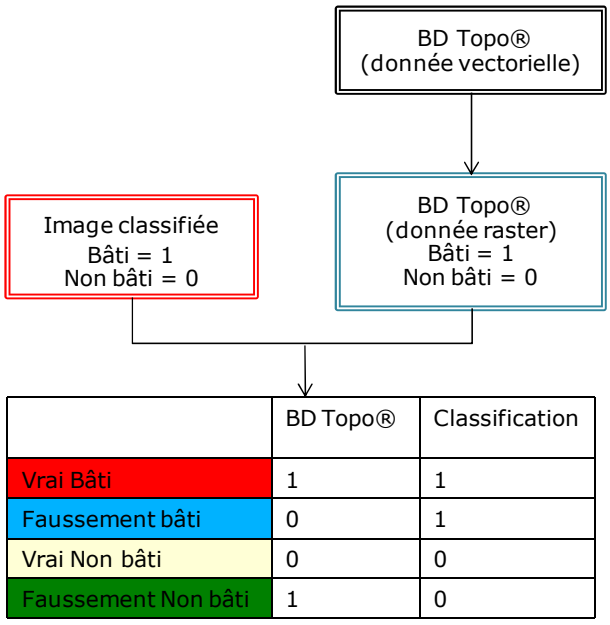
Figure 3-31 : Détails des échantillons classifiés par la régression logistique



3.3 Validation des classifications

Afin de comparer et de valider les extractions réalisées par les quatre méthodes de classification exposées auparavant, la BDTopo® (cf.2.3.2) a été utilisée comme source des données de référence. Pour ce faire, la BDTopo® a été rastérisée avec une taille de pixel de 50cm (taille des pixels des images), puis une matrice de confusion a été construite (fig. 3-32). Par la suite, la légende des couleurs adoptée est la suivante : la couleur rouge représente le « vrai bâti », c'est-à-dire les pixels bien classés comme bâti ; la couleur bleu symbolise le « faussement bâti », c'est-à-dire les pixels classés à tort comme bâti ; la couleur beige indique le « vrai non bâti », c'est-à-dire les pixels bien classés comme non bâti ; et la couleur vert identifie le « faussement non bâti », c'est-à-dire les pixels classés à tort comme non bâti.

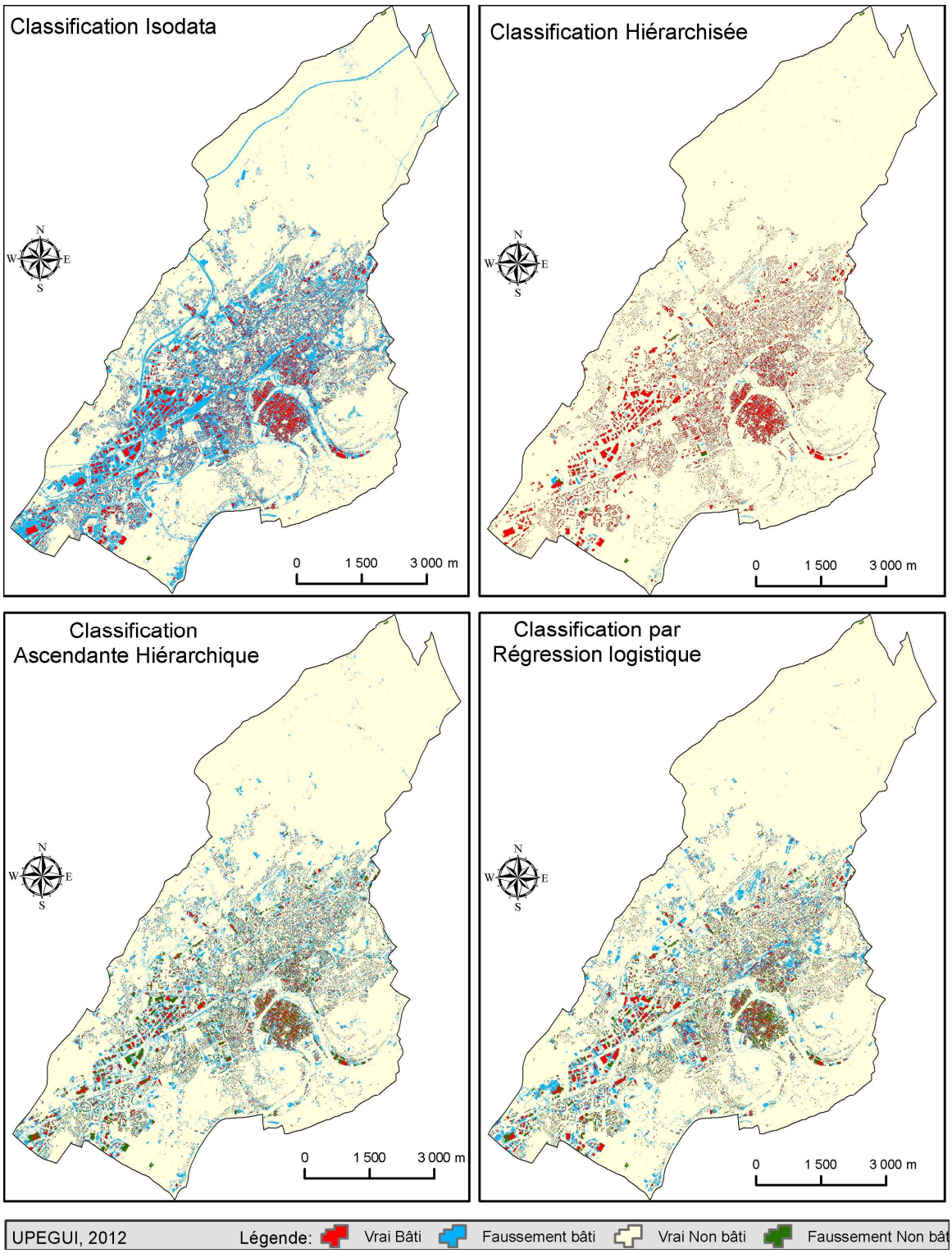
Figure 3-32 : Diagramme de la construction des matrices de confusion



	BD Topo®	Classification
Vrai Bâti	1	1
Faussement bâti	0	1
Vrai Non bâti	0	0
Faussement Non bâti	1	0

Ainsi une comparaison visuelle et globale peut se faire sur les quatre classifications (fig. 3-33). La classification ISODATA accumule le nombre le plus élevé des pixels « faussement bâti », où l'on peut identifier clairement le réseau routier. La classification hiérarchique est dominée par le « vrai bâti » et le « vrai non bâti ». Dans la classification ascendante hiérarchique prédomine le « faussement non bâti », *a priori* sur les immeubles de grande taille. Quant à la classification par régression logistique, de petites plages de pixels « faussement bâti » et « faussement non bâti » donnent un effet « poivre et sel » sur la grande tache du « vrai non bâti »

Figure 3-33 : Comparaison des quatre classifications par rapport au « bâti » de la BDTopo®

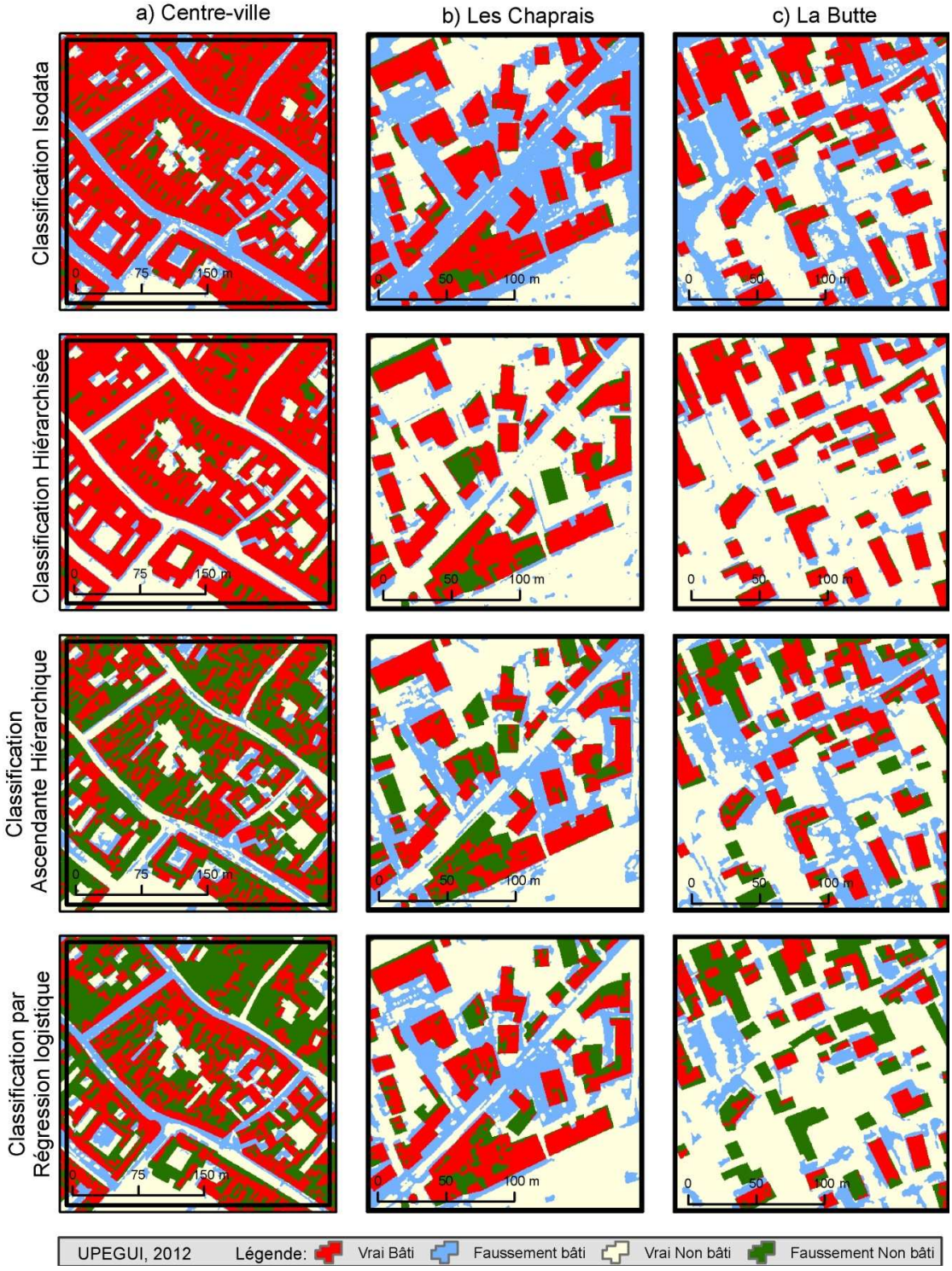


Pour analyser en détail les classifications sur chacun des échantillons par typologie de quartier, les échantillons ont été regroupés en 3. Ainsi, sur le premier groupe composé par le centre-ville et la première couronne, on peut déduire que :

- les pâtés de maisons identifiés par la BDTopo® font apparaître des cours intérieures dans « faussement non bâti » en sachant qu'elles sont des zones « non bâties ». Cela se voit clairement sur le centre-ville et sur les quatre classifications,
- la différence dans la date de mise à jour des données (BDTopo®2006 vs Images 2009) fait apparaître comme « bâti » des zones qui, à l'heure actuelle ne le sont pas, comme dans la cas des Chaprais où une zone démolie et pourtant bien classée dans la catégorie « non bâti » par la classification hiérarchisée apparait dans la validation des résultats comme « faussement non bâti »,
- les parties de la toiture qui sont recouvertes par la végétation, sont, elles classées -dans les quatre classifications- comme « faussement non bâti ». Cela peut se voir dans la partie nord-est de l'échantillon du centre-ville, ou encore dans la partie est de l'échantillon des Chaprais,
- un certain décalage est observé -malgré le géoréférencement des données- par exemple à La Butte, où les bâtiments présentent d'un côté une frange « faussement non bâti » et du côté opposé présentent une frange « faussement bâti ». L'origine de ce phénomène vient peut-être des différentes données utilisées comme sources pour « restituer[42] » les bâtiments. Il faut signaler que ce décalage n'est pas systématique,
- dans les trois échantillons, la classification ISODATA a extrait quasiment tous les « bâtis » ; et donc le nombre des pixels « faussement non bâti » est moindre que celui des pixels « faussement bâti »,
- dans les deux cas de classification fondée sur l'« objet », le nombre des pixels « faussement bâti » est moindre que celui des pixels « faussement non bâti ». En effet, les hypothèses de segmentation des bâtiments visent à approcher les toitures. C'est la raison pour laquelle on peut observer que les pâtés de maisons du centre-ville, ou encore certains bâtiments sur Les Chaprais ou sur La Butte, sont constitués par les différents pans des toitures, classés comme « vrai bâti » et « faussement non bâti ».

[42] « En photogrammétrie, processus d'obtention d'une représentation (graphique, numérique, photographique) à trois dimensions d'un objet à partir de clichés pris avec une chambre métrique » (CILF, 1997)

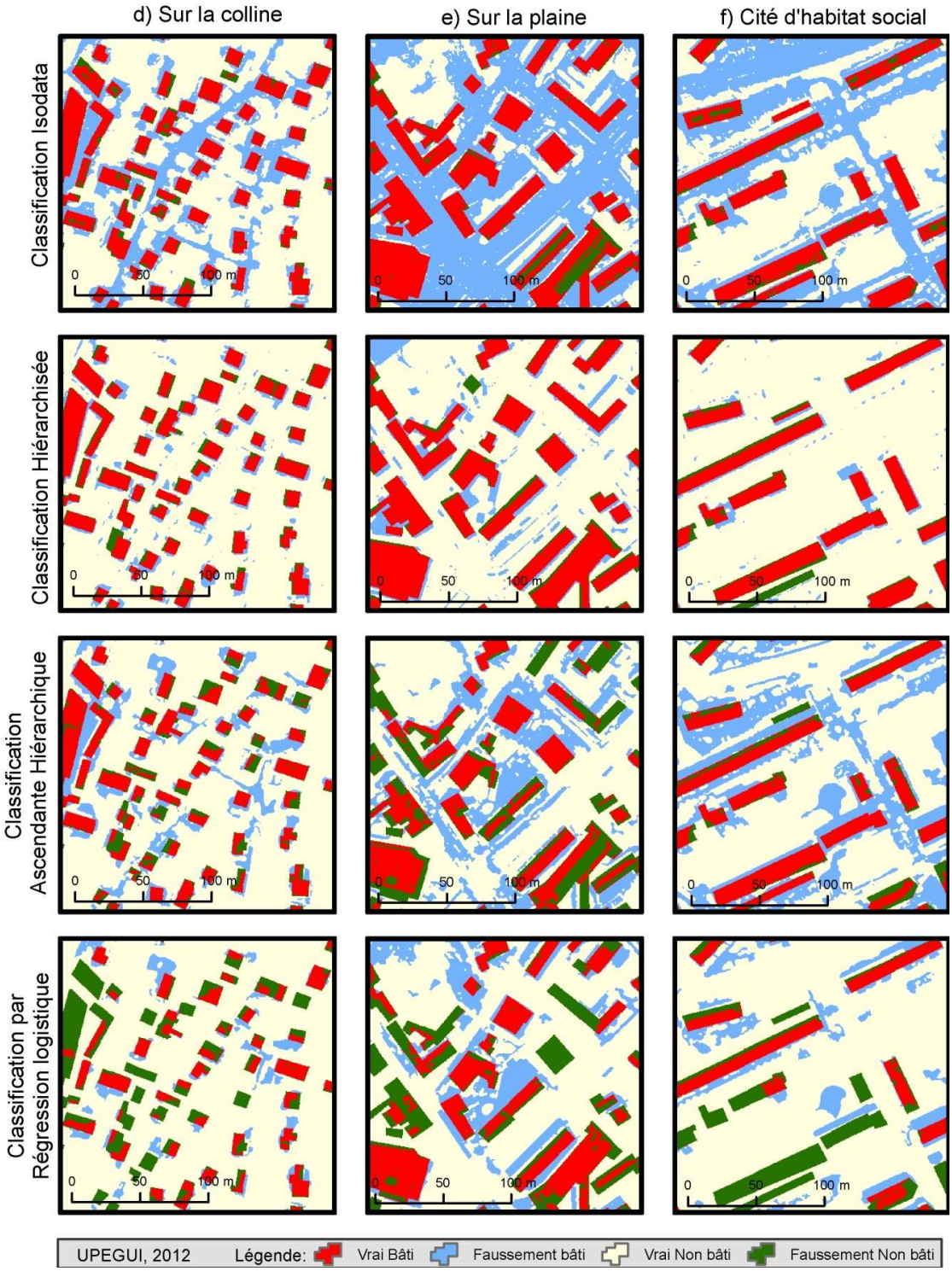
Figure 3-34 : Validation des résultats des quatre classifications sur les détails des échantillons, correspondant au centre-ville et à la première couronne



Quant aux échantillons de la deuxième couronne (fig. 3-35), la performance des classifications est similaire à celle de la première couronne ; cela malgré la variation de la taille des bâtis qui va de la petite taille (sur la colline), jusqu'à la grande taille en largeur (les cités d'habitat social), en passant par la taille moyenne (sur la plaine). En outre, on observe

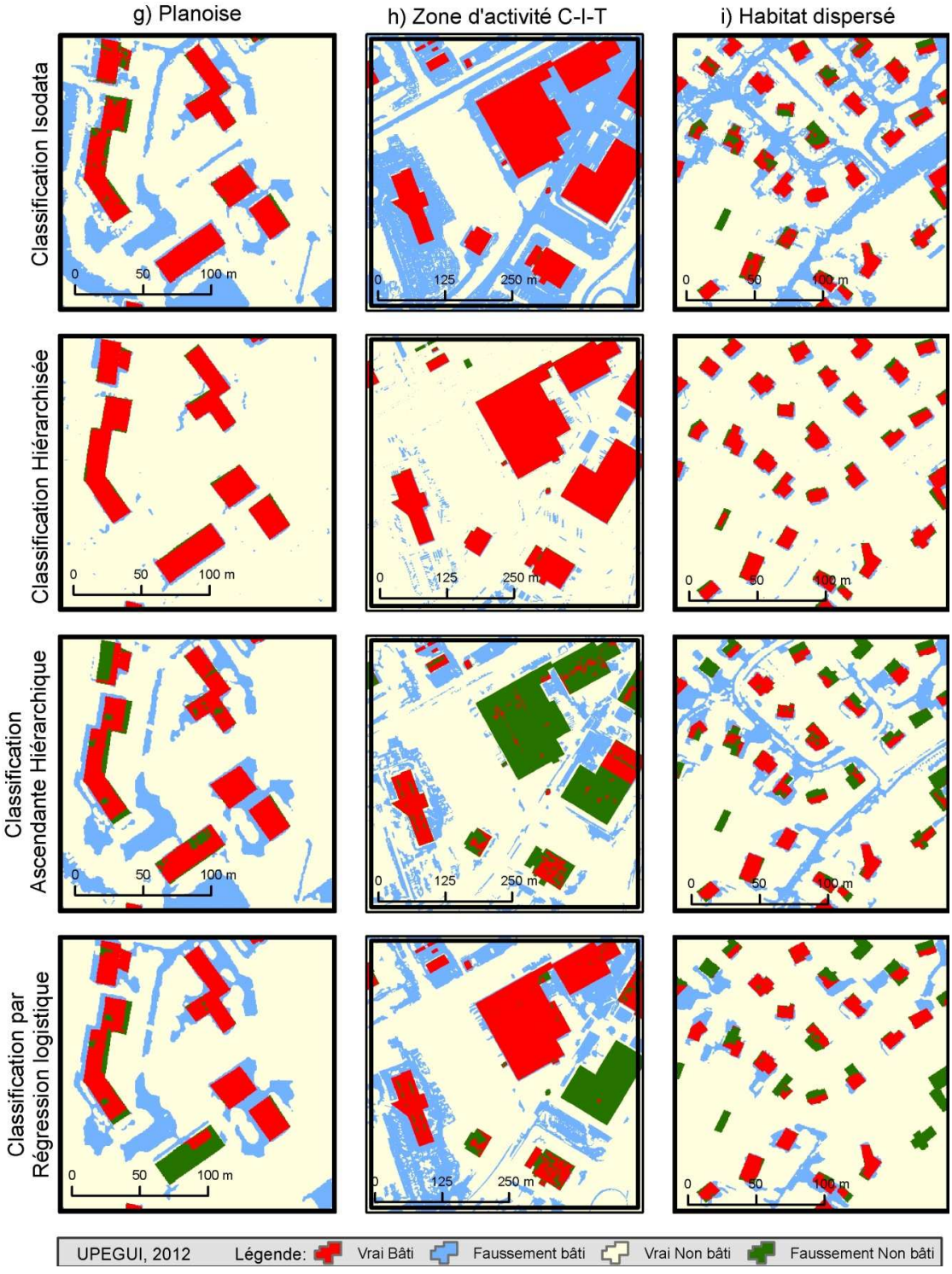
que le nombre de pixels « faussement bâti », correspondant aux routes, est plus importante dans les deux classifications non dirigées (classification ISODATA et classification ascendante hiérarchique) que dans les deux classifications dirigées (classification hiérarchisée et classification par régression logistique).

Figure 3-35 : Validation des résultats des quatre classifications sur les détails des échantillons, correspondant à la deuxième couronne



Pour les échantillons de la troisième couronne (fig. 3-36), on constate une performance des classifications similaire à celles des autres échantillons. En outre, on observe que les grands immeubles ont été mieux extraits par la classification par régression logistique que par la classification ascendante hiérarchique. Toutefois, il en résulte que les parkings sont classés dans « faussement bâti » dans la zone d'activité C-I-T.

Figure 3-36 : Validation des résultats des quatre classifications sur les détailles des échantillons, correspondant à la troisième couronne



Par ailleurs, pour compléter la précédente analyse visuelle -globale et détaillée- et pour opérer une validation quantitative des classifications, les matrices de confusion ont été construites, pixel-à-pixel (tab. 3-5), pour chacune des classifications. Ainsi, la précision globale (correspondant au pourcentage de pixels bien classés par rapport aux données de référence) et le coefficient kappa ont été calculés.

Tableau 3-5 : Précision de l'extraction du « bâti », comparaison des quatre méthodes

	Méthode de classification	**BD TOPO®** Bâti	**BD TOPO®** Non bâti	Précision Globale[43]	Coefficient kappa[44]
Approche par pixel	**ISODATA**				
	Bâti	17 725 036	41 623 880	0,83	0,37 p<10e-09
	Non bâti	2 378 343	199 360 562		
	Classification hiérarchisée				
	Bâti	17 856 479	5 280 211	0,97	0,81 p<10e-09
	Non bâti	2 246 900	235 704 231		
Approche par objet	**Classification ascendante hiérarchique**				
	Bâti	10 275 928	20 340 572	0,88	0,34 p<10e-09
	Non bâti	9 827 451	220 643 870		
	Régression logistique				
	Bâti	10 684 898	17 411 461	0,90	0,39 p<10e-09
	Non bâti	9 418 481	223 572 981		

Différents enseignements peuvent être tirés de l'analyse de ces résultats.

Premier point, les valeurs de précision globale sont plus hautes (le nombre est plus grand) que celles du coefficient kappa, en raison : d'une part, du fait que le Kappa tient compte de l'intervention du hasard ; d'autre part, de la proportion des pixels « non bâti » par rapport à celle des pixels « bâti » : à savoir 240 984 442 pixels (92,27% de la surface totale de la ville) vs 20 103 379 pixels (7,73% de la surface totale de la ville).

[43] Précision globale = $Po = \left(\frac{vrai\ bâti + vrai\ non\ bâti}{nb\ total\ des\ pixels}\right)$

[44] Coefficient kappa = K= $\frac{Po-Pe}{1-Pe}$, où *Po* est la proportion d'accord observée (précision globale), et *Pe* est la proportion d'accord aléatoire.

Second point, toutes les classifications présentent une bonne précision globale, supérieure au 0,8. Par ordre décroissant, la première est celle obtenue par la classification hiérarchisée avec 0,97 ; la deuxième est celle de la classification par régression logistique, avec 0,90 ; la troisième correspond à la classification ascendante hiérarchique avec 0,88 ; et la dernière est celle de la classification ISODATA avec 0,83.

Par ailleurs, toutes les valeurs du coefficient kappa sont significatives statistiquement, p<10e-09.

De plus, la valeur du kappa de la classification hiérarchisée atteint 0,81, selon l'interprétation du coefficient kappa proposée par Landis et Koch (1977) : il présente un accord « excellent » entre cette classification et les données de référence. C'est-à-dire que cette classification offre la meilleure répartition « bâti » / « non bâti ».

Pour finir, les trois autres classifications présentent un accord « faible » (kappa variant entre 0,21 et 0,40) avec les données de référence : la valeur kappa obtenue par la classification par régression logistique correspond à 0,39 ; celle de la classification ISODATA atteint 0,37 ; et celle de la classification ascendante hiérarchique abouti 0,34. Toutefois, la classification par régression logistique tend plutôt à un accord « modérée » (0,41 -0,60).

En synthèse, après analyse de toutes les classifications, sous l'aspect méthodologique et du point de vue des résultats, cinq conclusions peuvent être tirées :

En premier lieu, les classifications non dirigées, soit dans une approche par pixel soit dans une approche par objet, surestiment les « bâtiments », bien que l'on soit plus proche du bâti avec l'approche par objet.

En deuxième lieu, la classification ISODATA est une technique non dirigée fréquemment utilisée, mais elle est très dépendante de l'opérateur. Son application aux données THRS est lourde en raison du temps de calcul. Elle permet d'identifier les zones imperméables ; autrement dit elle permet une discrimination des zones végétales du reste. En revanche, elle ne permet pas d'approcher les « bâtiments ». D'ailleurs, des résultats semblables ont été obtenus par Lu et son équipe (2010).

En troisième lieu, la classification hiérarchisée est facilement compréhensible par l'utilisateur et son application est assez rapide, en termes de temps de calcul et de la taille du stockage. Néanmoins, la reproductibilité de cette méthode est limitée en raison de son recours aux données altimétriques. Cela rend difficile son accessibilité pour deux raisons majeures. La première : le coût de ces données est très élevé[45], limitant ainsi son accès aux pays en voie de développement. La deuxième raison : elles sont aéroportées, ce qui signifie qu'il faut faire des campagnes de vol, qui requièrent des autorisations de l'Etat, parfois difficiles à obtenir.

En quatrième lieu, la classification fondée sur l'objet est une alternative au recours à des données altimétriques. Cependant, le temps de calcul est plus important dans cette approche, de même que la taille du stockage. En effet, en partant d'une image en teinte de gris, au moins 52 nouvelles images ont été créées dans la première partie de la segmentation. De plus, 18 images classifiées ont été nécessaires pour identifier les bâtiments potentiels et ainsi achever la segmentation. En sus, si l'on opte pour l'approche dirigée (régression logistique dans notre cas) on doit avoir recours aux zones d'apprentissage (autour de 13 000 objets dans notre application) qui demandent : d'une part, du temps disponible de la part de l'opérateur ; et d'autre part, son expertise pour l'identification des bâtiments. Néanmoins, cette méthode est plus déterministe, et donc plus reproductible que celle de la classification par ISODATA. En outre, on peut déduire des résultats obtenus (qualitativement et quantitativement) que les attributs morphologiques sont nécessaires mais non suffisants pour identifier les « bâtiments » avec la démarche méthodologique par morphologie mathématique que nous avons utilisée. Au final, notre approche « hybride » de la classification fondée sur l'objet -avancée dans cette thèse-, donne de bons résultats.

En dernier lieu, la classification ascendante hiérarchique évite l'étape d'apprentissage nécessaire dans une classification dirigée, bien qu'elle soit un peu moins performante. De plus elle peut s'appliquer à des échantillons de petite taille.

[45] « Acquisition de données Lidar sur 220 km2 dans la région de Franche-Comté (secteurs de Besançon et Mandeure-Mathay, Doubs) financement Région de Franche-Comté /MSHE C.N. Ledoux, projet Lieppec, 127 613 euros) » (Nuninger et Ostir, 2009)

Nous avons présenté dans ce chapitre quatre méthodes différentes de classifications que nous avons utilisées afin d'identifier « bâti / non bâti ». Une grande partie du temps d'exécution de cette recherche a été investie dans cette étape d'extraction des bâtiments parce qu'elle est liée à l'un des côtés novateurs de notre recherche : l'utilisation des données THRS qui offre une large gamme des scenarios possibles au moment de reproduire les méthodes (de l'image en teinte de grise, à l'utilisation des données altimétriques issues du données LiDAR, en passant par les images multispectrales). Ainsi, nous avons appliqué des approches fondées sur le pixel : d'une part, la méthode classique de la classification non dirigée (ISODATA) ; et d'autre part, la classification dirigée (classification hiérarchisée). Nous avons également mis au point et avons implémenté de nouvelles méthodes de classification, notamment celles fondées sur l'objet. De plus, étant donné la configuration de notre zone d'étude, c'est-à-dire une zone recouverte par deux images différentes, les méthodes proposées ont été testées et validées en dehors des données utilisées pour le mettre au point. De même, ces méthodes ont été analysées dans une optique de santé publique (prévision et action sanitaire) car c'est le cadre de notre étude et l'objectif majeur de cette recherche. En outre, une validation qualitative et quantitative a été faite afin de préciser la portée de chacune des classifications proposées. Au total, toutes les méthodes sont utilisées par la suite.

« L'art ou le jeu de la modélisation consistent à rechercher par un choix conscient et raisonné, par une expérimentation guidée par des hypothèses, comment on peut représenter de la façon plus efficace, c'est-à-dire la plus fidèle et la plus puissante à la fois, une organisation spatiale »
R. Brunet (La composition des modèles dans l'analyse spatiale in Espace géographique, 1980)

Chapitre 4. Estimation des populations : de la surface au volume

L'estimation de la population repose sur la variable « bâti / non bâti », obtenue dans l'étape précédente (cf. chapitre 3). Pour ce faire, deux indicateurs et deux approches de modélisation différents sont utilisés. Les estimations des populations modélisées sont ensuite validées à travers les effectifs de population fournis par l'INSEE, permettant, d'une part, d'expérimenter la méthodologie proposée, et d'autre part de pallier à l'une des limites fondamentales de la modélisation de la population, à savoir l'impossibilité de valider exactement les résultats obtenus (Weber, 2005).

En ce qui concerne le découpage « administratif » de l'INSEE (1999) pour le recensement de la population de 1999, deux « unités géographiques de base » sont disponibles : l'IRIS et l'îlot. Besançon compte, d'un côté, 52 IRIS (surface variant entre 0,11km² et 17km², ce qui correspond à une étendue -avec un pixel de 0,5m- de 459 923 pixels à 67 996 763 pixels) ; et de l'autre côté, 680 îlots (surface variant entre 0,0009km² et 17km², correspondant à une étendue -avec un pixel de 0,5m- de 3 694 pixels à 67 996 763 pixels). Les résultats obtenus dans cette étape de modélisation de la population, et donc de l'analyse spatiale, seront rapportés, tout au long du chapitre, au niveau des IRIS, et les résultats des îlots étant reportés en annexe 4[46]. Cela permet d'évaluer la performance de la méthode pour une unité géographique donnée.

[46] En effet, les îlots ne seront pas utilisés dans l'estimation de taux d'incidence, car le très faible nombre de nouveaux cas enregistrés par ilôt rend le calcul de taux d'incidence inopportun (à cause de leur très grande fluctuation aléatoire). Cependant, nous avons considéré important d'analyser cette unité géographique, car *a priori*, c'est l'échelle la plus adéquate pour étudier, du point de vue géographique, le niveau de détail des données THRS.

Une analyse exploratoire de la « structure spatiale » de la population a été initialement faite à travers deux paramètres, « la densité » et le « quotient de localisation ». Le « quotient de localisation » caractérise le degré de concentration d'une population (ou sous-population) dans une unité spatiale en la comparant à toutes les autres unités spatiales d'un même ensemble. Cette information relative à la taille des unités spatiales est utile et pertinente quand on s'intéresse aux effectifs de la population (Pumain et Saint Julien, 2010b).

La densité $\left(\frac{nombre\ d'habitants}{surface}\right)$ varie selon l'unité d'analyse (fig. 4-1). On voit ainsi apparaître sur les mêmes zones, par exemple au sud-est, des densités plus importantes dans les îlots, que dans les IRIS. Pour les IRIS, une densité moyenne et forte est observée dans les quartiers suivants : le centre ville ; la Butte ; les Chaprais ; Planoise ; sur la seconde couronne (plaine), notamment sur la partie plus au nord ; et sur la colline, mais seulement au nord-est. A l'échelle des îlots, on repère certaines cités d'habitat social, ainsi que d'autres zones denses sur la colline. Néanmoins, la densité ne prend pas en compte « l'effet de taille[47] » et c'est pour cela que l'on fait appel au « quotient de localisation » - Q_{ij}[48] (fig. 4-2). Q_{ij} peut être supérieur à 1 : la population est concentrée ; égal à 1 : il n'y a pas de concentration particulière de population ; ou inférieur à 1 : la population n'est pas concentrée. Ainsi, lorsqu'on prend en compte cet « effet de la taille » on observe que certaines unités spatiales (soit IRIS, soit îlot) à faible densité (par exemple le secteur ouest de la ville) deviennent « concentrées » dans le quotient de localisation ; ou l'inverse. Plus précisément, sur les IRIS le Q_{ij} met en évidence la concentration de la population à l'ouest de la ville : sur l'habitat dispersé, sur la colline, et sur la plaine. S'agissant des îlots, le Q_{ij} a un effet contraire : il présente quelques zones du centre ville et sur la colline faiblement concentrées.

[47] « On appelle *effet de taille* la perturbation introduite dans certaines comparaisons (éventuellement menées avec des calculs comme le coefficient de corrélation ou l'analyse en composantes principales) du fait de la grande ressemblance de l'ordre établi entre des unités spatiales à cause de leurs inégalités de dimension, pour des variables mesurées en effectives mais qui sont censées décrire autre chose que ces inégalités » (Pumain et Saint Julien, 2010)

[48] $Qij = \left(\frac{(Xij/X.j)}{(Xi./X..)}\right)$ où X_{ij} est le nombre d'habitants j dans l'unité spatiale i ; $X_{.j}$ correspond au nombre total d'habitants j dans l'ensemble des unités spatiales ; $X_{i.}$ est la surface de l'unité spatiale i ; et $X_{..}$ correspond à la surface totale dans l'ensemble des unités spatiales

Figure 4-1 : Densité de la population : à droite sur les îlots, à gauche sur les IRIS.

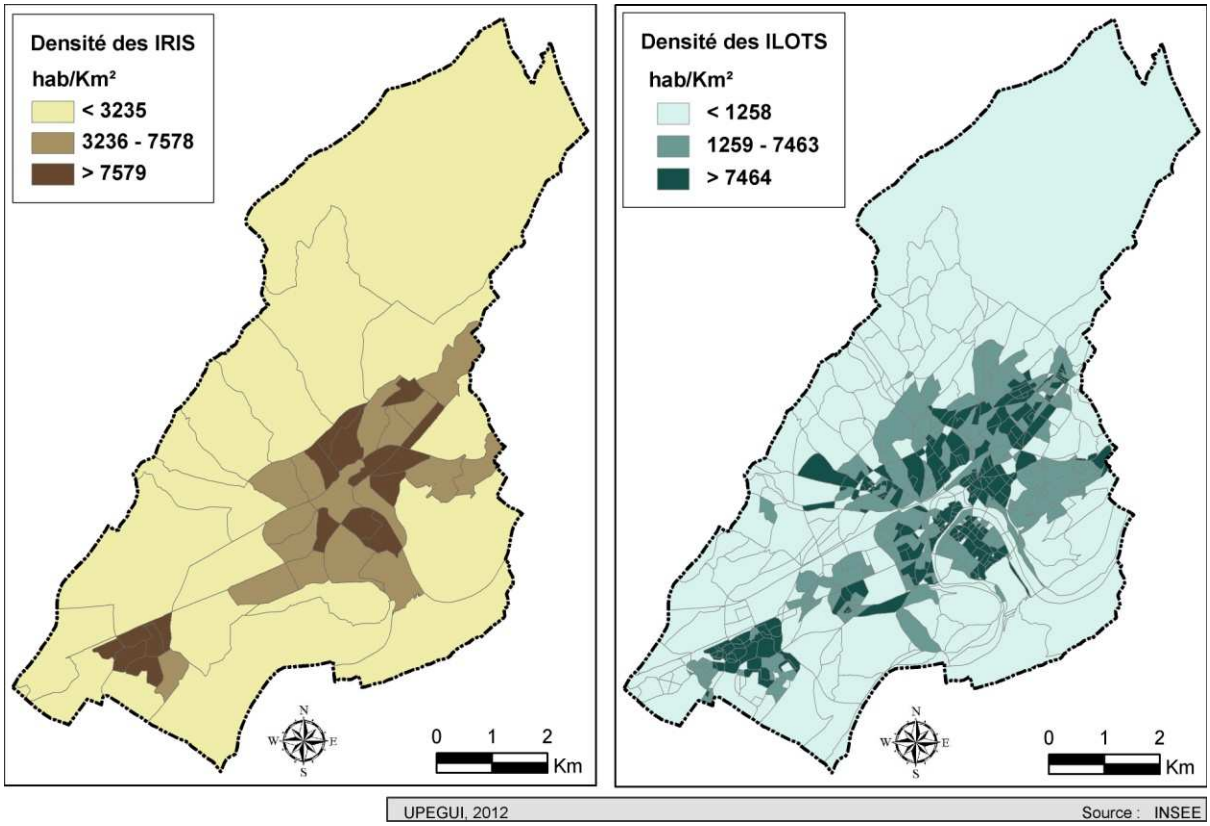
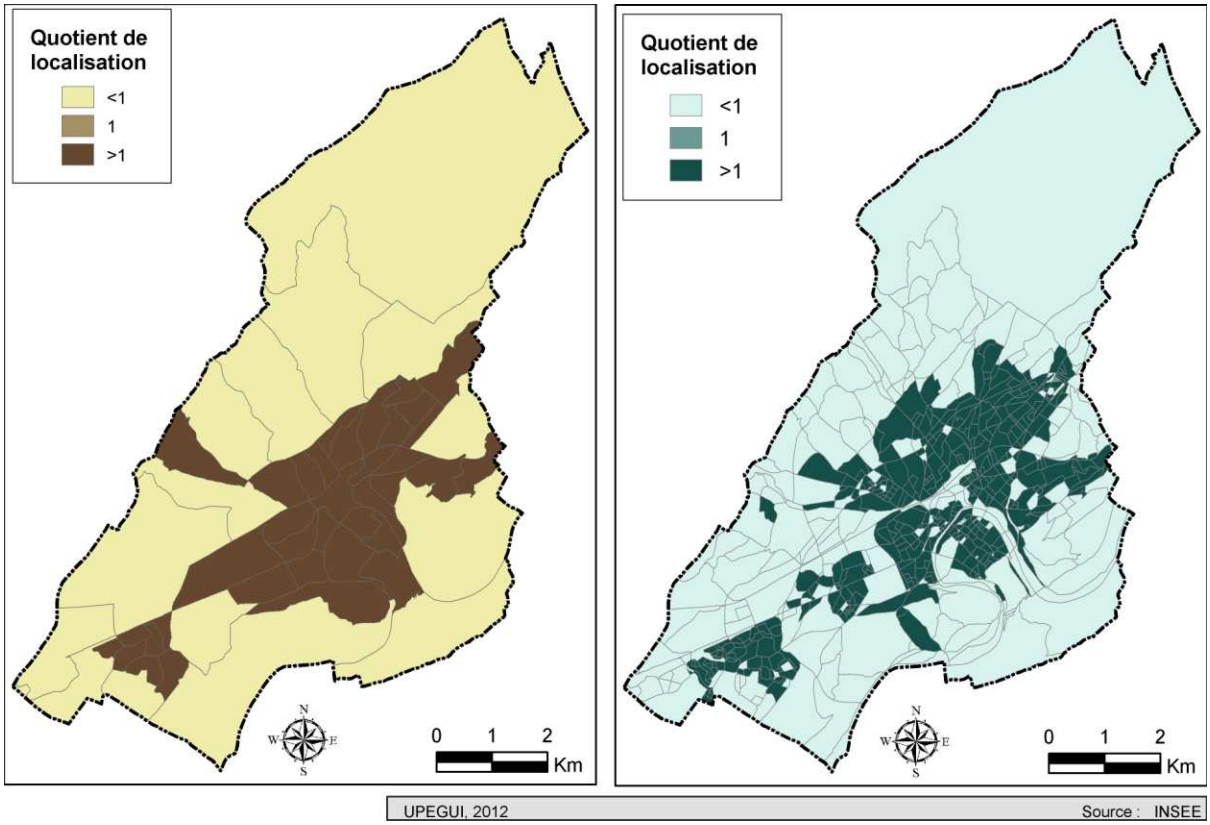


Figure 4-2 : Quotient de localisation de la population : à droite sur les îlots, à gauche sur les IRIS



4.1 Méthode dasymétrique et indicateurs de répartition

Selon Wu et son équipe (2008), les estimations de la population (ou de la distribution de la population) qui utilisent à la fois le recensement et les variables utiles pour définir la population, sont généralement les plus fiables. D'ailleurs, Fisher et Langford (1995) ont démontré une réduction de 54% d'erreur d'estimation de la population, en utilisant la méthode dasymétrique au lieu de « l'interpolation aréale » (c'est-à-dire de type surface - cf. 1.1.1) simple. C'est cette méthode dasymétrique (cf. 1.1.3) que nous avons retenue pour modéliser la distribution spatiale de la population.

Du point de vue théorique, MacEachren (1994) a placé les cartes dasymétriques dans le continuum entre les cartes isoplèthes et les cartes choroplèthes, avançant ainsi que les cartes dasymétriques représentent les données à mi-chemin entre les surfaces statistiques « lisses et échelonnées » (*smooth and stepped*, en anglais). Néanmoins, Eicher et Brewer (2001) signalent que le terme « dasymétrique » a, depuis le début de la recherche en « interpolation aréale », des limites ambiguës entre la « carte » et la « méthode ». Ainsi, la méthode dasymétrique décrit une forme particulière d'«interpolation aréale » car elle utilise des données auxiliaires : par exemple la méthode binaire, ou encore les autres méthodes « intelligentes ». En revanche, les cartographes emploient le terme « dasymétrique » pour se référer à un type général de carte thématique produite en utilisant un éventail de méthodes (Cuff et Mattson, 1982 ; Maciejewski, 2011 ; Longley et *al.*, 2011). Par ailleurs, Eicher et Brewer (2001) proposent cinq critères qui guident l'utilisation de la méthode dasymétrique, parmi lesquels les deux premiers permettent la différenciation des cartes dasymétriques des isoplèthes et/ou choroplèthes. Le premier critère permet de signaler l'homogénéité des zones, c'est-à-dire que les zones dasymétriques finales de la carte sont intérieurement homogènes. Le deuxième critère concerne le caractère abrupt des frontières entre zones, marquant les changements de la variable représentée. Le troisième critère est la désagrégation des données source. Le quatrième correspond à l'interpolation « intelligente », c'est-à-dire à une interpolation guidée par des données auxiliaires au lieu de l'utilisation simple du poids des surfaces. Le dernier critère considère « l'imbrication » des

données, c'est-à-dire l'interpolation des données à un niveau plus détaillé que celui de la surface totale.

De plus, Eicher et Brewer (2001) proposent une catégorisation de la méthode dasymétrique, en fonction de l'utilisation des données auxiliaires. La première catégorie correspond à la méthode de « la variable limitante » où la zone pour distribuer la population est réduite de manière arbitraire, en fonction des connaissances de l'expert. La deuxième catégorie, la méthode binaire : les données auxiliaires sont divisées en deux catégories, par exemple peuplé/non peuplé. Finalement, la dernière catégorie, la méthode « trois-classes », divise en trois catégories les données auxiliaires, en les pondérant (de manière subjective) pour distribuer la population.

La méthode dasymétrique est apparue en même temps que les autres méthodes d'interpolation aréale, mais son utilisation a été reportée pendant de nombreuses années (Eicher et Brewer, 2001). Cela s'explique par la difficulté inhérente à la construction de l'indicateur de désagrégation (ou répartition) (Langford, 2007), mais surtout par la difficulté à obtenir l'information pour le construire (Maantay *et al.*, 2007 ; Lwin et Murayama, 2011). La diffusion des systèmes d'information géographique (SIG) (Eicher et Brewer, 2001 ; Holt et Lu, 2011 ; Langford *et al,*. 2008 ; Mennis, 2003, Wu *et al*., 2005), associée à l'avènement de la collecte systématique des données de télédétection ont renouvelé l'intérêt pour cette méthode, offrant ainsi de nouvelles opportunités dans son application (Holt et Lu, 2011). En effet, les SIG constituent une aide à la conception de modèles par la facilité d'accès et de mise en œuvre de l'information sur l'espace étudié, devenant un outil puissant pour instrumenter une modélisation (Miellet, 1992).

En définitive, la modélisation (redistribution) spatiale de la population par la méthode dasymétrique infère deux procédures de transformation (fig. 4-3) : l'application dasymétrique qui consiste en la désagrégation des données ; et l'interpolation spatiale aréale qui, à son tour, réagrège les données dans une unité géographique donnée (Holt et Lu, 2011). La désagrégation spatiale suppose que les données fournies globalement pour toute une région peuvent être redistribuées dans cette région à travers un paramètre local. Cela implique l'utilisation de données auxiliaires liées à la variation/distribution de la population, afin de donner un sens à la répartition de la population, et donc de rendre plus significative son résultat (fig. 4-4). Un choix judicieux des différentes échelles d'analyses est dès lors crucial dans ce type de modélisation, à savoir : l'échelle de l'unité géographique de

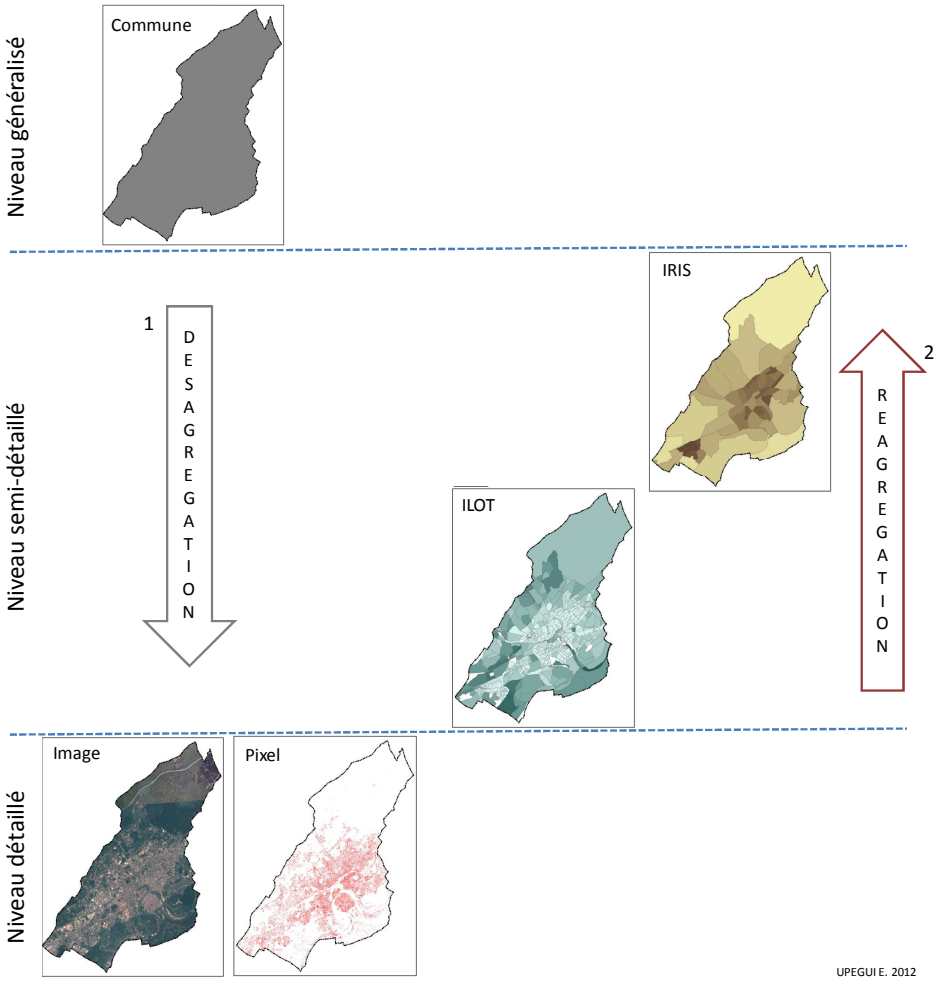
« départ » avec la population totale (niveau d'analyse général) ; l'échelle du paramètre local qui permet la désagrégation (niveau d'analyse détaillé) ; et l'échelle de l'unité géographique « d'arrivée » souhaitée pour faire la réagrégation (fig. 4-3). Si l'on ramène ces niveaux d'analyse aux échelles des produits cartographiques de l'IGN et à notre problématique, on peut dire que :

- l'échelle de départ coïncide avec la limite administrative « Commune », laquelle fait partie de la BDCarto® produite à une échelle qui varie entre 1 :100 000 et 1 :50 000 ;

- l'échelle du paramètre local -construit à partir des données télédétection à 50cm de résolution spatiale- qui permet la désagrégation, peut s'aligner sur celle de la BDOrtho® qui varie entre 1 :5 000 et 1 :2 000 ;

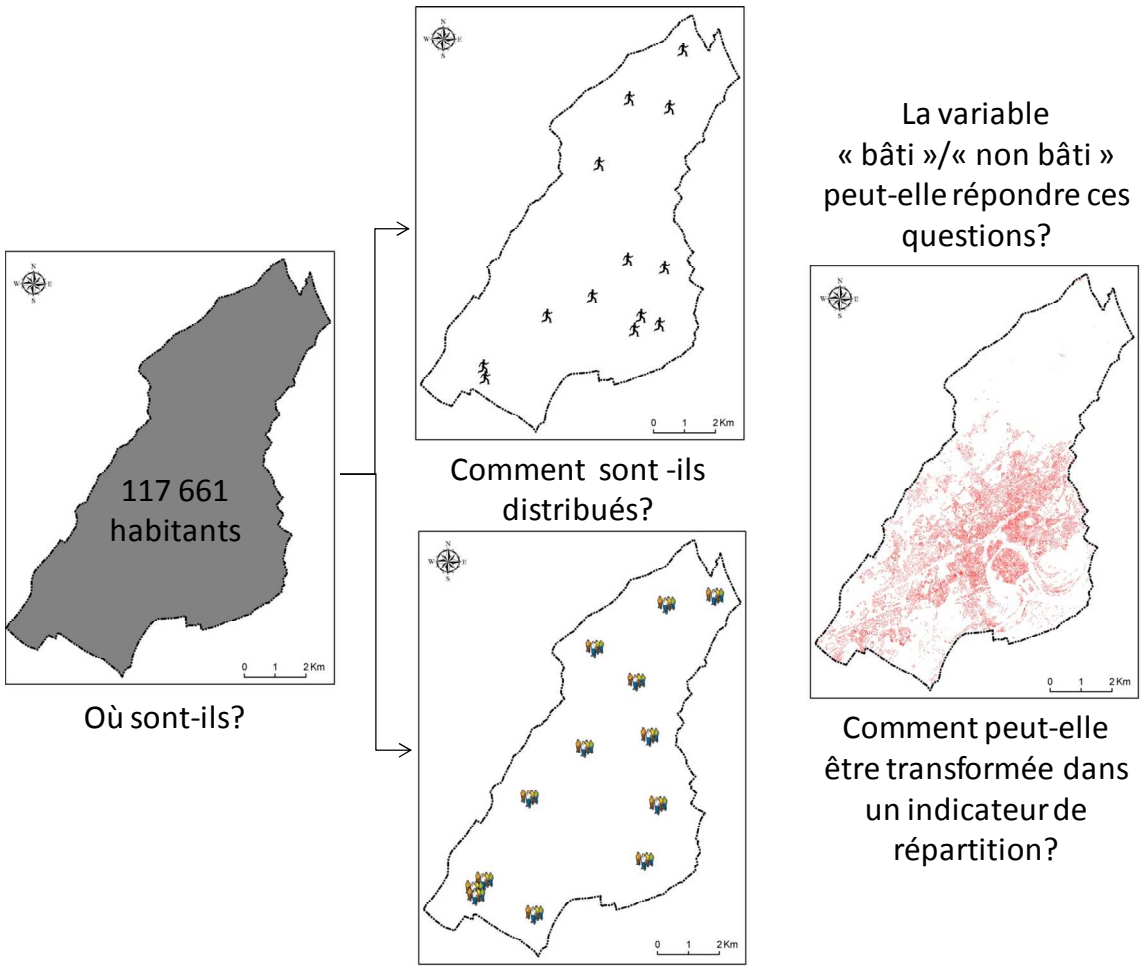
- l'échelle d'arrivée correspond à une échelle infra-communale qui peut se comparer aussi bien aux produits Scan 25® -à échelle 1 :25 000- que BDTopo® -exploitable à des échelles allant du 1 :5 000 au 1 :50 000-. Cette échelle d'arrivée est représentée dans cette étude par les unités géographiques de l'IRIS et de l'îlot.

Figure 4-3 : Illustration du principe de la méthode dasymétrique



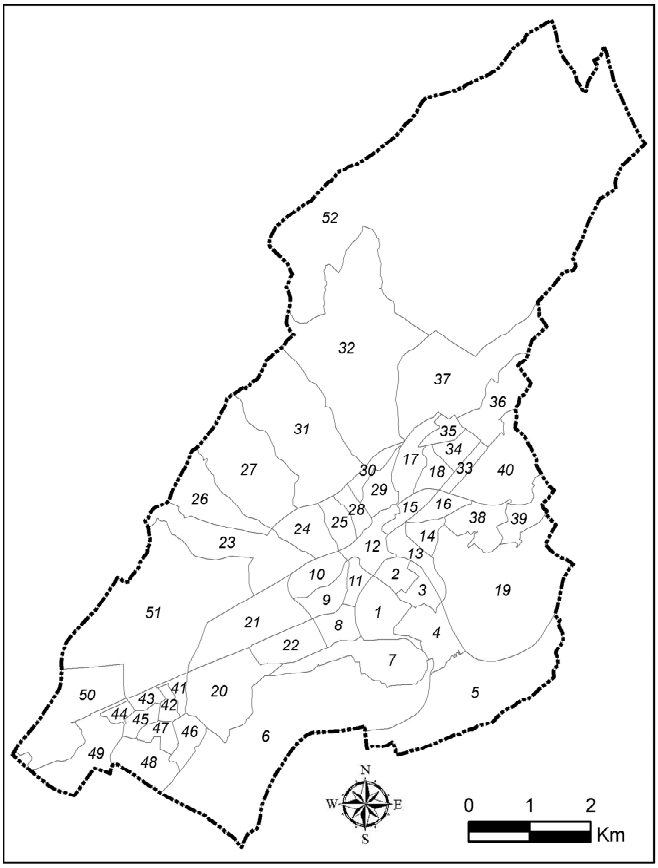
En pratique, en raison des objectifs de santé publique (prévision et action sanitaire) de cette étude (où l'estimation de la population doit être: simple, automatique, reproductible, et exportable), le focus est réalisé dans le cadre de la catégorie « binaire » de la méthode dasymétrique. Ainsi, à partir de la variable choisie (« bâti »/ « non bâti ») nous avons construit d'une part, les indicateurs de répartition (paramètre local) qui permettent la désagrégation de la population ; et d'autre part, l'algorithme mathématique décrivant la relation entre la population et ces indicateurs. Ces indicateurs restent « simples » dans notre cas, bien qu'ils puissent être complexifiés à loisir, avec par exemple les statistiques de semi-variance de la texture d'une image (Wu *et al.*, 2006). Aussi, comme cette recherche se veut en continuité avec les travaux de Viel et Tran (2009), nous partons des indicateurs de répartition proposés par eux afin de comparer nos résultats aux leurs. Ce choix découle de notre objectif secondaire qui est d'évaluer l'apport des données THRS *vs* HRS dans le même cadre urbain.

Figure 4-4 : Le rôle des données auxiliaires dans la désagrégation de la population



Par soucis de clarté, nous avons recours à la dénomination des IRIS (fig. 4-5) pour analyser les résultats obtenus ; même si nous continuons à faire allusion aux différents quartiers (cf. chapitre 2). Enfin, toutes les cartes sont représentées graphiquement par leur distribution en quartiles, même si les légendes (les intervalles de la population) ne sont pas identiques par définition.

Figure 4-5 : Identification et localisation des IRIS sur la commune de Besançon



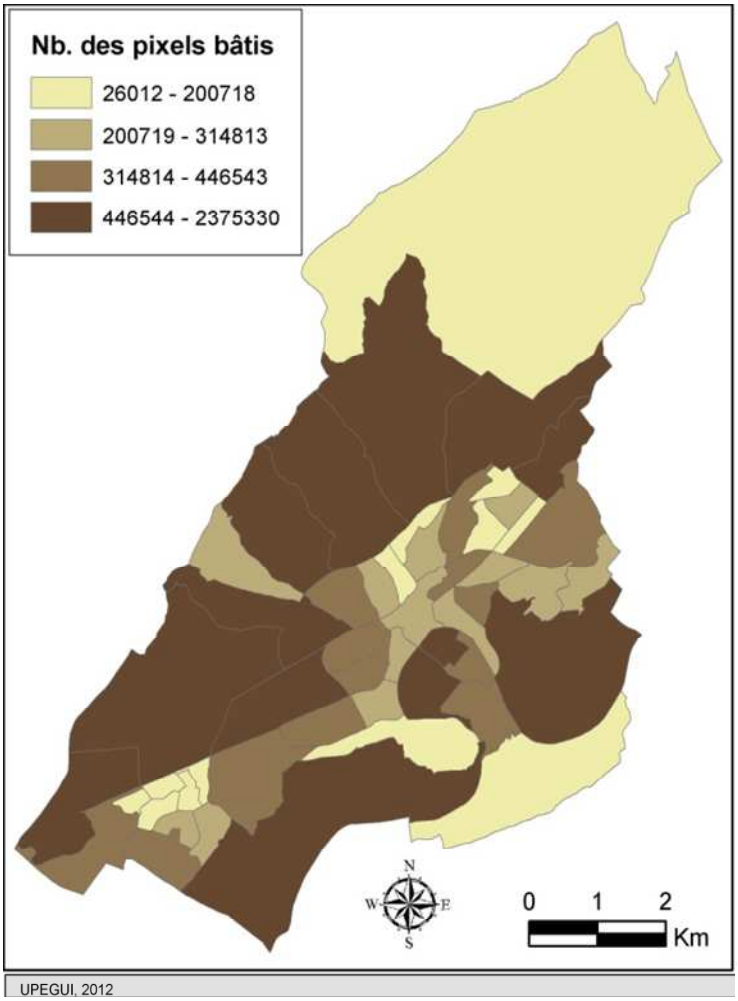
ID	NOM IRIS	ID	NOM IRIS
1	CHAMARS	27	MONTBOUCONS
2	REPUBLIQUE	28	MONTJOUX
3	SARRAIL	29	VIOTTE
4	CITADELLE	30	TREY
5	CHAPELLE DES BUIS	31	RUE DE VESOUL
6	VELOTTE	32	TORCOLS
7	GRETTE	33	RUE DE BELFORT
8	VIEILLES PERRIERES	34	CRAS
9	VILLARCEAU	35	CHOPIN
10	XAVIER-MARMIER	36	PALENTE
11	MARULAZ	37	SARAGOSSE
12	RUE BATTANT	38	CHAFFANJON
13	MOUILLERE	39	CLAIRS-SOLEILS
14	FONTAINE-ARGENT	40	VAITE
15	ROTONDE	41	EPOISSES-CHAMPAGNE
16	CHAPRAIS	42	EPOISSES-BOURGOGNE
17	CHASNOT	43	ILE-DE-FRANCE
18	PARC DES CHAPRAIS	44	PIEMONT
19	BREGILLE	45	DIDEROT
20	ROSEMONT	46	VICTOR HUGO
21	SAINT-FERJEUX	47	CASSIN
22	RISLER	48	LAFAYETTE
23	OBSERVATOIRE	49	SAINT-LAURENT
24	MONTRAPON	50	CHATEAUFARINE
25	FONTAINE-ECU	51	TILLEROYES
26	BOULOIE	52	CHAILLUZ

UPEGUI, 2012 Source: INSEE

4.1.1 Pixel-bâti : la somme des pixels

La manière plus simple de transformer la variable « bâti »/« non bâti » en un indicateur de répartition est celle du dénombrement des pixels par unité géographique. Ces unités (IRIS) sont pondérées, dans la modélisation de la population, par leur nombre de pixels (cette méthode correspond à la méthode NUM dans le travail de Viel et Tran, 2009). Ainsi, en guise de référence, les « bâtis » de l'IGN ont été utilisés pour spatialiser l'indicateur (fig. 4-6), car ces bâtis ont constitué la « vérité terrain » pour valider l'extraction des bâtiments (cf. chapitre 3). En outre, cette référence est appliquée dans tous les calculs autour de la modélisation de la population.

Figure 4-6 : Indicateur pixel-bâti selon les bâtis de l'IGN



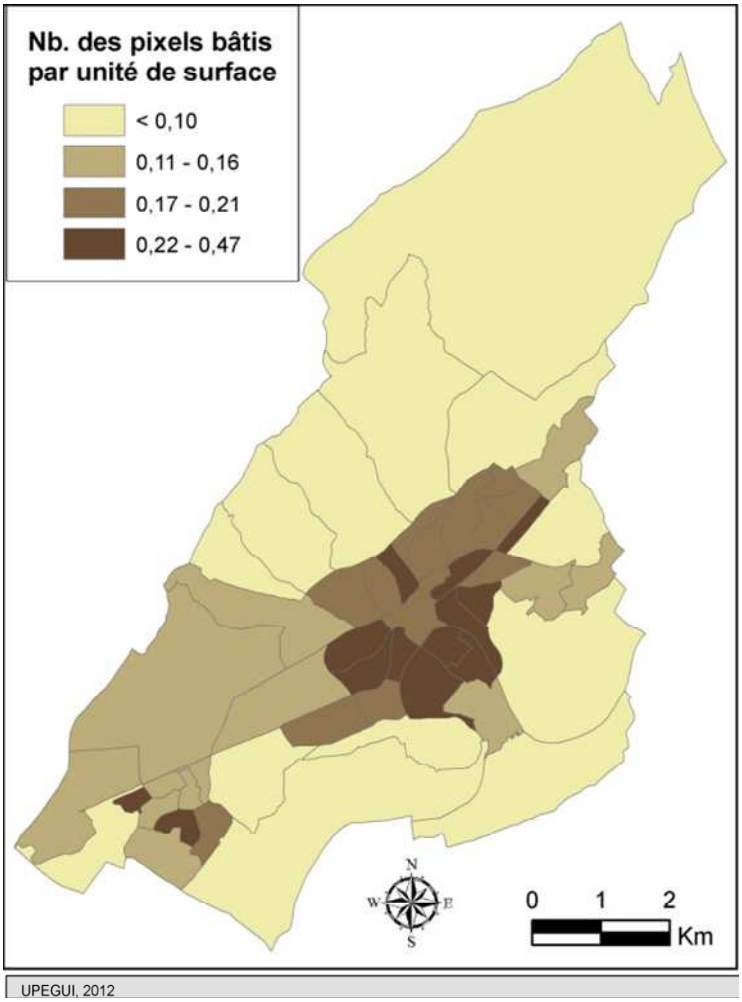
Avec cet indicateur, la plupart des IRIS, de la dernière enveloppe de la ville, se localisent dans le dernier quartile -le plus grand nombre des pixels bâtis- de la distribution des pixels. Fait exception l'IRIS « Chailluz » (nommé ainsi en référence à la forêt de Chailluz), laquelle compose une seule unité géographique (tant IRIS qu'îlot), qui se situe tout au nord de la ville, et qui fait partie du premier quartile -le plus petit nombre des pixels bâtis- de la distribution des pixels. En analysant la structure spatiale, on observe quasiment une « inversion » par rapport aux unités géographiques (fig. 4-2) qui concentrent la population.

4.1.2 Densité du bâti : le nombre de pixels par unité de surface

Une autre manière simple de transformer la variable « bâti »/ « non bâti » en indicateur de répartition touche à la densité. Celle-ci permet de « raisonner » la quantité de pixels « bâtis » par rapport à la surface, facilitant la comparaison entre différentes unités géographiques. Ainsi les IRIS sont pondérées, dans la modélisation de la population, par leur

densité (cette méthode correspond à la méthode PDI dans le travail de Viel et Tran, 2009). De la même manière que dans l'indicateur précédent (pixel bâti), les « bâtis » de l'IGN ont été utilisés pour spatialiser l'indicateur (fig. 4-7).

Figure 4-7 : Indicateur de la densité du bâti selon les bâtis de l'IGN



Cet indicateur recoupe pratiquement la structure spatiale de la densité de la population illustrée dans la figure 4-1. Toutefois, la représentation diffère selon les quantiles employés : la première (fig. 4-1) a été représentée par terciles et la seconde (fig. 4-7) a été représentée par quartiles. Au-delà de cette précision, les IRIS correspondant à la zone d'activité C-I-T se différencient clairement dans le deuxième quartile -densité du bâti faible- de la distribution des pixels.

4.2 Approches 2D -3D

Pour l'étape de la modélisation, une fois les indicateurs de répartition définis (pixel-bâti et densité du bâti), il est nécessaire d'établir une équation mathématique. Cet algorithme de répartition doit satisfaire à trois conditions nécessaires pour distribuer la population (Briggs *et al*., 2007) : la première est que les populations estimées doivent être non-négatives ; la deuxième est que les populations doivent être non-saisonnières, c'est-à-dire qu'elles doivent être stables dans le temps et dans l'espace. Enfin, la dernière condition est liée à la propriété « pycnophylactique » (conservation du volume) proposée par Tobler (1979), c'est-à-dire que la population modélisée ne doit pas dépasser le paramètre global, à savoir la population totale.

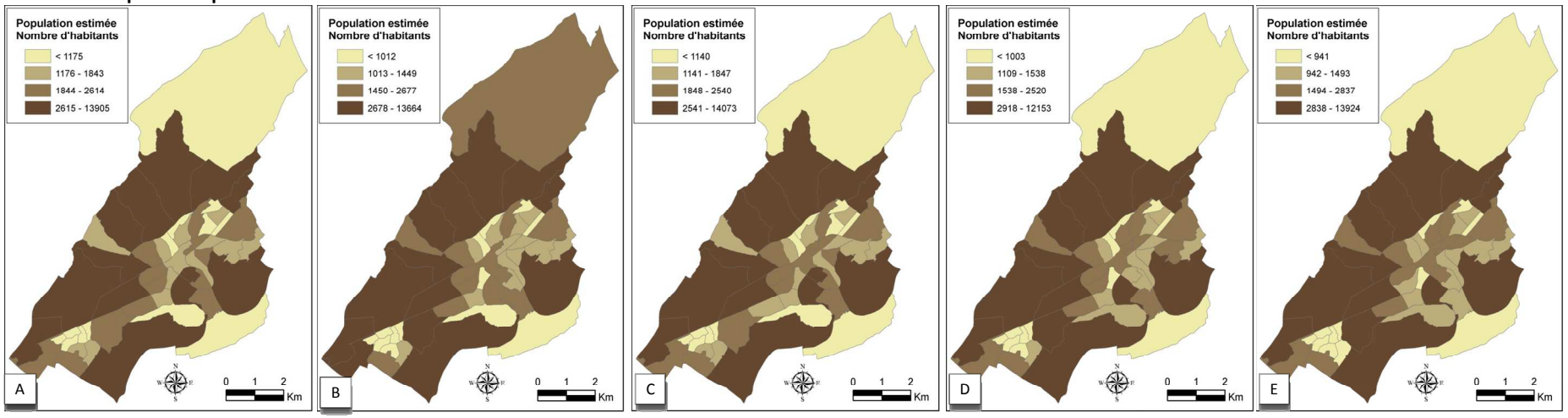
En utilisant les deux indicateurs de répartition (pixel-bâti et densité du pixel), deux approches de modélisation « surface » et « volume » sont proposées.

4.2.1 L'approche « surface »

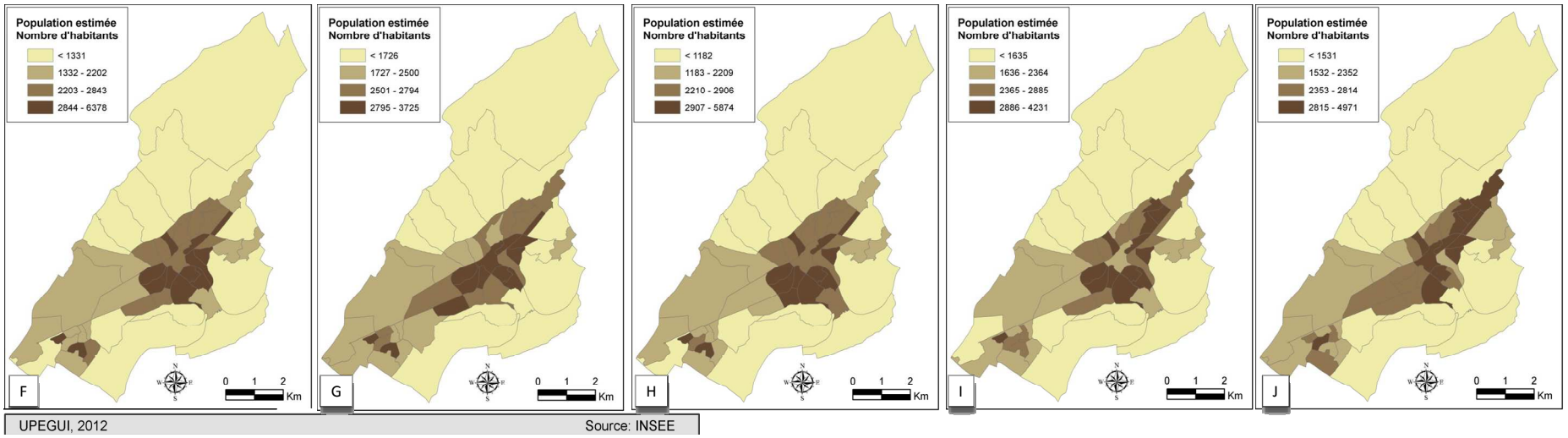
Nous sommes partie des deux méthodes proposées par Viel et Tran (2009), et nous les avons regroupées dans l'approche « surface », laquelle se fonde sur la surface des images classifiées (« bâti »/ « non bâti ») dans l'étape précédente (cf. chapitre 3). Au total, pour chacune des deux méthodes (pixel-bâti et densité du bâti) cinq populations ont été estimées (fig. 4-8 de A à E par la première méthode ; de F à J par la deuxième méthode) : d'un côté, chacune des quatre populations par chacune des quatre classifications ; de l'autre, la cinquième par les bâtis de l'IGN ayant servi comme données de référence.

Figure 4-8 : Population estimée par l'approche surface : de A à E par la méthode MPBP, de F à J par la méthode MDBP

Méthode du pixel bâti pondérée



Méthode de la densité du bâti pondérée



4.2.1.1 Méthode du pixel-bâti pondérée

La première méthode, que nous appellerons « méthode du pixel-bâti pondérée » (MPBP), utilise la somme des pixels (voir 4.1.1) comme pondération dans la modélisation (équation 1).

Équation 1. Méthode du pixel-bâti pondérée

$$popi = \frac{pop * \sum_{j=1}^{ni} pij}{\sum_{i=1}^{N} \sum_{j=1}^{ni} pij}$$

Où pop_i est la population estimée pour l'unité géographique *i* (soit l'IRIS soit l'îlot) ; *pop* est la population totale de la commune ; p_{ij} est la valeur du pixel *j* dans l'unité géographique *i* ; n_i est le nombre de pixels inclus dans l'unité géographique *i* ; et *N* est le nombre total d'unités géographiques (à savoir : N=52 pour les IRIS, et N=680 pour les îlots).

Les résultats obtenus dans cette modélisation apparaissent dans la figure 4-8 : a) avec le « bâti » de l'IGN ; b) avec le « bâti » extrait par ISODATA ; c) avec le « bâti » extrait par classification hiérarchisée ; d) avec le « bâti » extrait par classification ascendante hiérarchique ; et e) avec le « bâti » extrait par régression logistique. Nous observons une certaine similitude entre les résultats. Ainsi, on trouve, globalement, dans le premier quartile (le moins peuplé) un groupe d'IRIS : à Planoise (Epoisses-Champagne, Epoisses-Bourgogne, Ile de France, Piémont, et Diderot), à la Chapelle des Buis, à la Grette et à Chailluz. Dans le dernier quartile (le plus peuplé), se localisent : les quartiers les plus périphériques (Saint-Ferjeux, Observatoire, Montboucons, Rue de Vesoul, Torcols, Palente, Saragosse) y compris les zones commerciales et industrielles (Chateaufarine, et Tilleroyes) ; certains quartiers du centre-ville (Chamars et République) ; ainsi que certains sur la deuxième couronne sur la colline (Bregille et Velotte). Les deux autres quartiles (deuxième et troisième) ont une distribution plus aléatoire et irrégulière selon le bâti employé pour la modélisation. C'est pour cela qu'on observe que Chailluz se situe dans le troisième quartile en raison de la surestimation des bâtiments résultant de l'extraction par ISODATA (voir 3.1.1) (fig. 4-8 b).

4.2.1.2 Méthode de la densité du bâti pondérée

La deuxième méthode, ci-après appelée « méthode de la densité du bâti pondérée » (MDBP), utilise la densité (voir 4.1.2) comme pondération (équation 2).

Équation 2. Méthode de la densité du bâti pondérée

$$popi = \frac{pop * Di}{\sum_{i=1}^{N} Di} \text{ où } Di = \frac{\sum_{j=1}^{ni} pij}{ni}$$

Où D_i est la densité de pixels de l'unité géographique *i ;* p_{ij} est la valeur du pixel *j* dans l'unité géographique *i ;* n_i est le nombre de pixels inclus dans l'unité géographique *i ;* pop_i est la population estimée pour l'unité géographique *i ; pop* est la population totale de la commune *;* et *N* est le nombre total d'unités géographiques (à savoir : N=52 pour les IRIS, et N=680 pour les îlots).

La figure 4-8 (de F à J) regroupe les résultats obtenus dans cette modélisation. La distribution sur la carte est la suivante : f) pour le « bâti » de l'IGN ; g) pour « bâti » extrait par ISODATA ; h) pour le « bâti » extrait par classification hiérarchisée ; i) pour le « bâti » extrait par classification ascendante hiérarchique ; et j) pour le « bâti » extrait par régression logistique). En analysant ces cartes, une inversion de la distribution spatiale s'observe par rapport à la méthode MPBP. Cependant, nous observons deux points en commun : d'une part, les résultats se ressemblent ; et d'autre part, la distribution varie selon le bâti employé pour la modélisation. Ainsi, globalement, les IRIS du premier quartile se localisent plutôt à l'extérieur de la ville, tandis que ceux du dernier quartile se retrouvent vers le centre-ville. La continuité de l'enveloppe extérieure est interrompue par les IRIS du deuxième quartile, notamment par ceux situés dans la zone d'activité C-I-T, à Planoise, et quelques-uns sur la deuxième couronne. Les IRIS du troisième quartile se localisent quasiment tous dans la deuxième couronne sur la plaine. De plus, les principaux axes routiers qui traversent la ville du sud-ouest au nord-est semblent influencer fortement la distribution

4.2.2 L'approche « volume »

Afin d'avancer dans la modélisation de la population, et d'exploiter de façon optimale les données dont nous disposons, nous avons proposé une nouvelle approche : l'approche « volume ». Cette approche se fonde également sur la surface des images classifiées

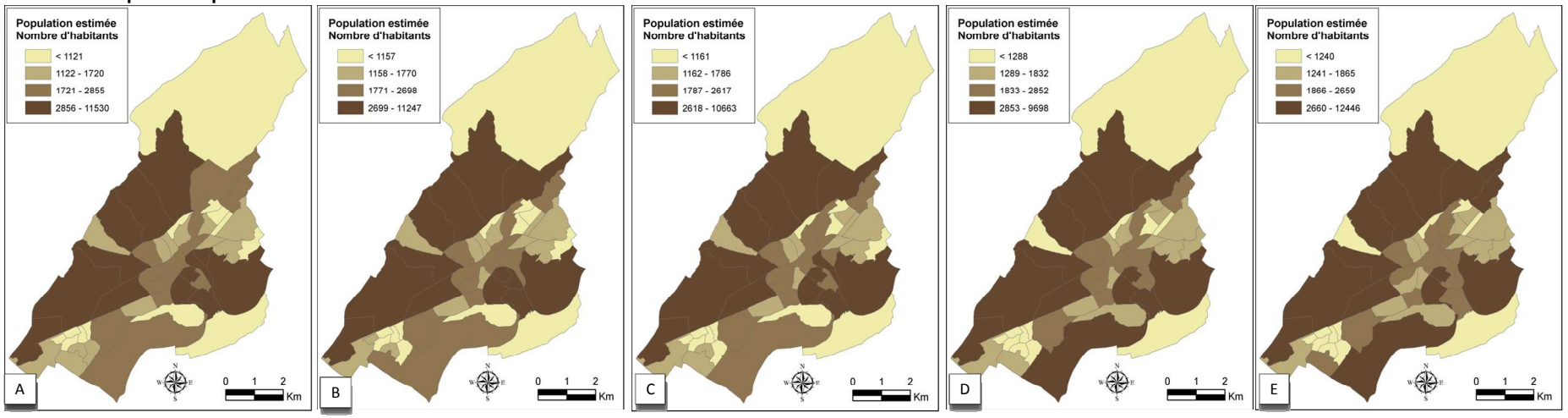
(« bâti »/ « non bâti ») dans l'étape précédente, mais en ajoutant comme critère la hauteur, issue des données LiDAR. Inclure la hauteur -dans la modélisation de la population- a pour objectif principal de différencier les maisons individuelles des immeubles collectifs, afin de leur attribuer un poids correct dans l'estimation de la population. En outre, cela permet de réduire l'effet des immeubles à utilisation commerciale / industrielle, qui, pour la plupart, correspondent à des grandes surfaces de plain pied, conduisant à une surestimation de la population dans une approche « surface ». De surcroît, si on prend en compte la hauteur, l'effet des pixels « faussement bâtis » sera également réduit : habituellement, les routes, les parkings, les places, etc. ne présentent pas d'élévation. Rappelons que la variable « hauteur » n'est pas facilement accessible et que pose une difficulté dans son obtention pour les diverses raisons exposées.

En pratique, nous avons utilisé la hauteur ($Hauteur = MNS - MNT$) calculée dans l'étape précédente (voir 3.2.1 - fig. 3-11g), sauf pour les données de référence de l'IGN (qui possèdent leur propre hauteur). Cette hauteur a été introduite dans le modèle, ainsi : à chaque pixel $_j$, dans l'unité de recensement $_i$, qui est affecté par une valeur p_{ij}, la valeur p_{ij} est égale à 0 si le pixel est « non bâti », où la valeur p_{ij} est égale à h_{ij} (h_{ij} étant la hauteur) si le pixel est « bâti ». La valeur de h_{ij} remplace alors la valeur 1, utilisée dans l'approche surface pour les pixels bâtis.

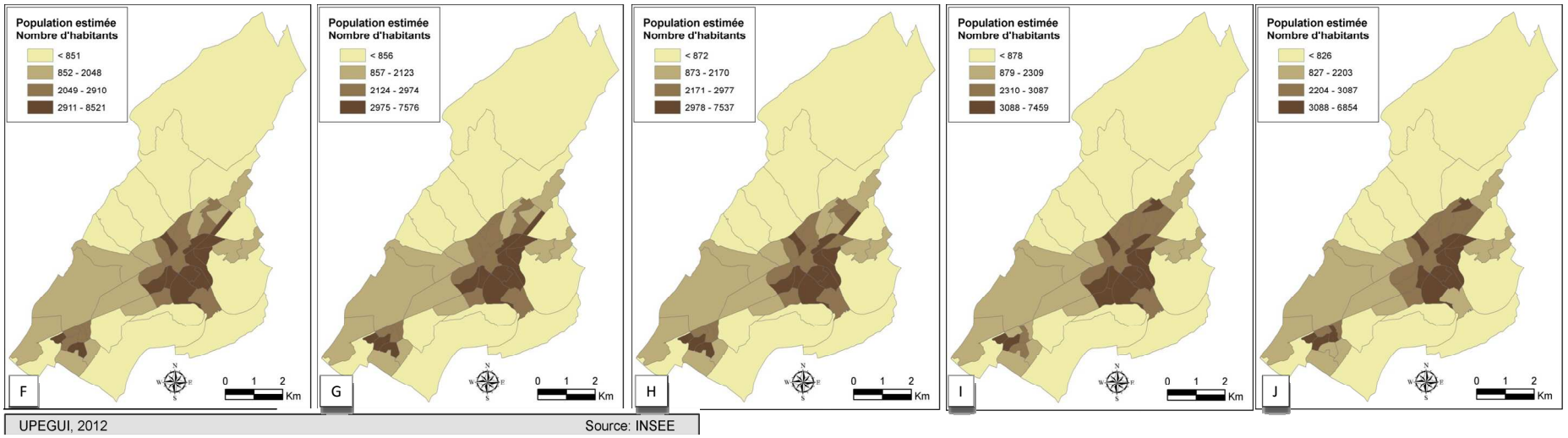
Enfin, de la même manière que dans l'approche « surface », cinq populations ont été estimées par chacune des deux méthodes (voir fig. 4-9 de A à E pour le pixel-bâti, et fig. 4-9 de F à J pour la densité du bâti) : d'un côté, une pour chacune des quatre classifications (voir chapitre 3) ; de l'autre, une pour le bâti de l'IGN ayant servi comme données de référence dans l'étape précédente.

Figure 4-9 : Population estimée par l'approche volume. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D

Méthode du pixel bâti pondérée en 3D



Méthode de la densité du bâti pondérée en 3D



4.2.2.1 Méthode du pixel-bâti pondérée en 3D

Reprenons l'équation 1 de la « méthode du pixel-bâti pondérée » (voir 4.2.1.1), utilisant la somme des pixels (voir 4.1.1). Les valeurs du pixel p_{ij}, correspondent maintenant aux volumes à la place des surfaces. Afin de différencier l'utilisation de la hauteur dans cette formule, cette méthode sera appelée ci-après « méthode du pixel bâti pondérée en 3D » (MPBP-3D).

Les résultats de cette modélisation volumétrique se retrouvent sur la figure 4-9 et correspondent d'une part, a) au « bâti » et la hauteur de l'IGN ; et d'autre part, aux bâtis couplés à la hauteur issue de données LiDAR, ainsi : b) « bâti » extrait par ISODATA ; c) « bâti » issu de la classification hiérarchisée ; d) « bâti » produit par classification ascendante hiérarchique ; et e) « bâti » extrait par régression logistique. Comme dans l'approche surface, la distribution spatiale de la population fait apparaitre des ressemblances entre les différents types de bâti, et la variation de cette distribution dépend du bâti modélisé. Globalement, le dernier quartile se localise en deux parties de la ville : d'une part, sur le centre-ville -excepté Sarrail- et vers l'est (Chamars, République, Citadelle, Mouillère, Fontaine-Argent et Bregille) ; d'autre part, vers l'extérieur de la ville (Saint-Ferjeux, Observatoire, Chateaufarine, Tilleroyes, Montboucons, Rue de Vesoul et, Torcols). Le premier quartile est éparpillé aux extrémités de la ville : au nord avec Chailluz ; au sud avec la Chapelle des Buis ; à l'est avec Clairs-Soleil ; et à l'ouest avec certains IRIS à Planoise. Les autres quartiles se distribuent aléatoirement sans concentration particulière, à l'exception du troisième quartile, qui se superpose aux principaux axes routiers.

4.2.2.2 Méthode de la densité du bâti pondérée en 3D

Nous qualifierons la méthode de la densité du bâti pondérée en approche volumétrique de « méthode de la densité du bâti pondérée en 3D » (MDBP-3D).

La figure 4-9 (de F à J) illustre les résultats obtenus dans cette modélisation. Les cartes son distribuées ainsi : f) pour le « bâti » et la hauteur de l'IGN ; g) pour le « bâti » extrait par ISODATA et la hauteur issue de données LiDAR ; h) pour le « bâti » extrait par classification hiérarchisée ainsi que la hauteur issue des données LiDAR ; i) pour le « bâti » résultant de la classification ascendante hiérarchique, et couplé à la hauteur (issue des données LiDAR) ; et

j) pour le « bâti » extrait par régression logistique couplé à la hauteur produite grâce au LiDAR. Comme déjà observé pour les autres modélisations, il existe une ressemblance entre les résultats. La distribution spatiale de la population estimée -utilisant les IRIS comme unité géographique- est plutôt régulière et concentrée sur toute l'étendue de la ville. Les IRIS du premier quartile se localisent : d'une part, sur la quasi-totalité de la deuxième couronne sur la colline ; et d'autre part, vers l'extérieur de la ville sur les quartiers en habitat dispersé. Le deuxième quartile se focalise surtout dans la zone d'activité C-I-T, tandis que le troisième quartile se situe sur Planoise, vers la Butte, ainsi que sur la deuxième couronne sur la plaine. Le dernier quartile est quasiment concentré au centre-ville et aux Chaprais.

4.3 Validation des estimations

4.3.1 Analyse des résultats

Afin de comparer et de valider les résultats obtenus avec les modélisations de la population (deux indicateurs de répartition -pixel-bâti et densité du bâti-, deux approches -surface et volume-, quatre méthodes de classification, ainsi que les données de référence de l'IGN - cf. chapitre 2), les données du recensement (IRIS et îlot, voir annexe 4 pour l'îlot) ont été utilisées comme données de référence. Pour ce faire, le coefficient de corrélation intra-classe[49] (CCI) a été calculé pour chacune des 20 estimations réalisées. Ces résultats apparaissent dans le tableau 4-1.

On constate que les estimations, ayant utilisé la « densité du bâti » comme indicateur de répartition, présentent des valeurs du CCI (entre 0,24 et 0,35) supérieures aux valeurs obtenues par les estimations qui ont utilisé le « pixel-bâti » comme indicateur de répartition (entre 0,05 et 0,13). D'autre part, toutes les estimations faites avec l'indicateur de répartition « densité du bâti » sont statistiquement significatives ($p < 0,05$), tandis que les estimations faites avec l'indicateur de répartition « pixel-bâti » ne le sont pas ($p > 0,05$).

Par ailleurs, l'utilisation des données de hauteur (c'est-à-dire l'approche « volume ») a amélioré les valeurs du CCI pour l'indicateur de répartition « pixel-bâti » (passant d'un intervalle de valeurs [-0,05 - 0,03] à un intervalle de valeurs [0,05 - 0,13]), mais ces données

[49] La principale différence entre ce coefficient de corrélation intra-classe et le coefficient interclasse (Pearson) repose sur le regroupement des données pour estimer leur moyenne et leur variance, cela afin de les centrer et de les mettre à la même échelle.

n'ont pas produit le même effet sur les valeurs du CCI pour l'indicateur « densité du bâti ». Ces dernières restent quasiment « stables » tant dans l'approche « surface » que dans l'approche « volume ».

De plus, les estimations de la population utilisant la classification hiérarchisée et la classification ascendante hiérarchique atteignent les valeurs les plus élevées, aussi bien dans les deux approches (surface et volume) que dans les deux méthodes de désagrégation (pixel-bâti et densité du bâti). Il est à noter que ces valeurs sont meilleures que les valeurs obtenues avec les données de référence du bâti de l'IGN.

Tableau 4-1 : Coefficient de corrélation intra-classe pour les IRIS : N=52

Méthode de modélisation de la population		**Méthode de désagrégation**			
		Pixel-bâti pondéré CCI (IC95%)	**p**	**Densité du bâti pondérée** CCI (IC95%)	**p**
Approche par surface	Bâti IGN	0,02 (-0,25-0,29)	0,44	0,26 (-0,00-0,50)	0,03
	C. ISODATA	-0,05 (-0,31-0,23)	0,63	0,26 (-0,00-0,50)	0,03
	C. Hiérarchisée	0,03 (-0,24-0,30)	0,42	0,29 (0,02-0,52)	0,02
	C. Ascendante Hiérarchique	0,03 (-0,24-0,30)	0,41	0,35 (0,09-0,57)	0,004
	C. Régression logistique	-0,01 (-0,28-0,26)	0,53	0,24 (-0,03-0,48)	0,04
Approche par volume	Bâti IGN	0,09 (-0,19-0,35)	0,27	0,25 (-0,02-0,49)	0,03
	C. ISODATA	0,08 (-0,20-0,34)	0,29	0,27 (-0,00-0,50)	0,02
	C. Hiérarchisée	0,09 (-0,18-0,36)	0,25	0,28 (0,01-0,51)	0,02
	C. Ascendante Hiérarchique	0,13 (-0,14-0,39)	0,17	0,29 (0,02-0,52)	0,02
	C. Régression logistique	0,05 (-0,22-0,32)	0,35	0,28 (0,00-0,51)	0,02

Chacune des estimations est confrontée aux données de recensement à l'aide d'un diagramme bi-varié. D'après l'analyse de ces figures (fig. 4-10 pour l'approche « surface » et fig. 4-11 pour l'approche « volume »), on peut déduire deux éléments.

En premier lieu, la MPBP (méthode du pixel bâti pondérée), tant par l'approche « surface » que par l'approche « volume » (fig. 4-10 de A à E, et fig. 4-11 idem), présente deux points (entourés d'un cercle pointillé dans les figures) qui exercent une forte influence sur la répartition de la population. La localisation de ces points sur le graphique indique une surestimation de la population dans le modèle. Ces deux points correspondent aux IRIS des Tilleroyes et de Châteaufarine, où l'utilisation des sols est majoritairement industrielle ou commerciale.

En deuxième lieu, la MDBP (méthode de la densité du bâti pondérée), tant par l'approche « surface » que par l'approche « volume » (fig. 4-10 de F à J, et fig. 4-11 idem), présente une distribution plus homogène que celle de la MPBP. Néanmoins, si on se réfère à la diagonale, on observe une inversion de la distribution des points des deux derniers quartiles par rapport aux deux premiers quartiles. Cela indique une sous-estimation des deux premiers quartiles (c'est-à-dire les moins peuplés) et donc une surestimation des deux derniers quartiles (les plus peuplés). Les résultats obtenus par cette modélisation ne s'ajustent donc qu'imparfaitement à la réalité de la distribution de la population.

Figure 4-10 : Diagrammes bi-variés des populations estimées par l'approche surface *vs* la population des IRIS1999. De A à E par la méthode MPBP ; de F à J par la méthode MDBP

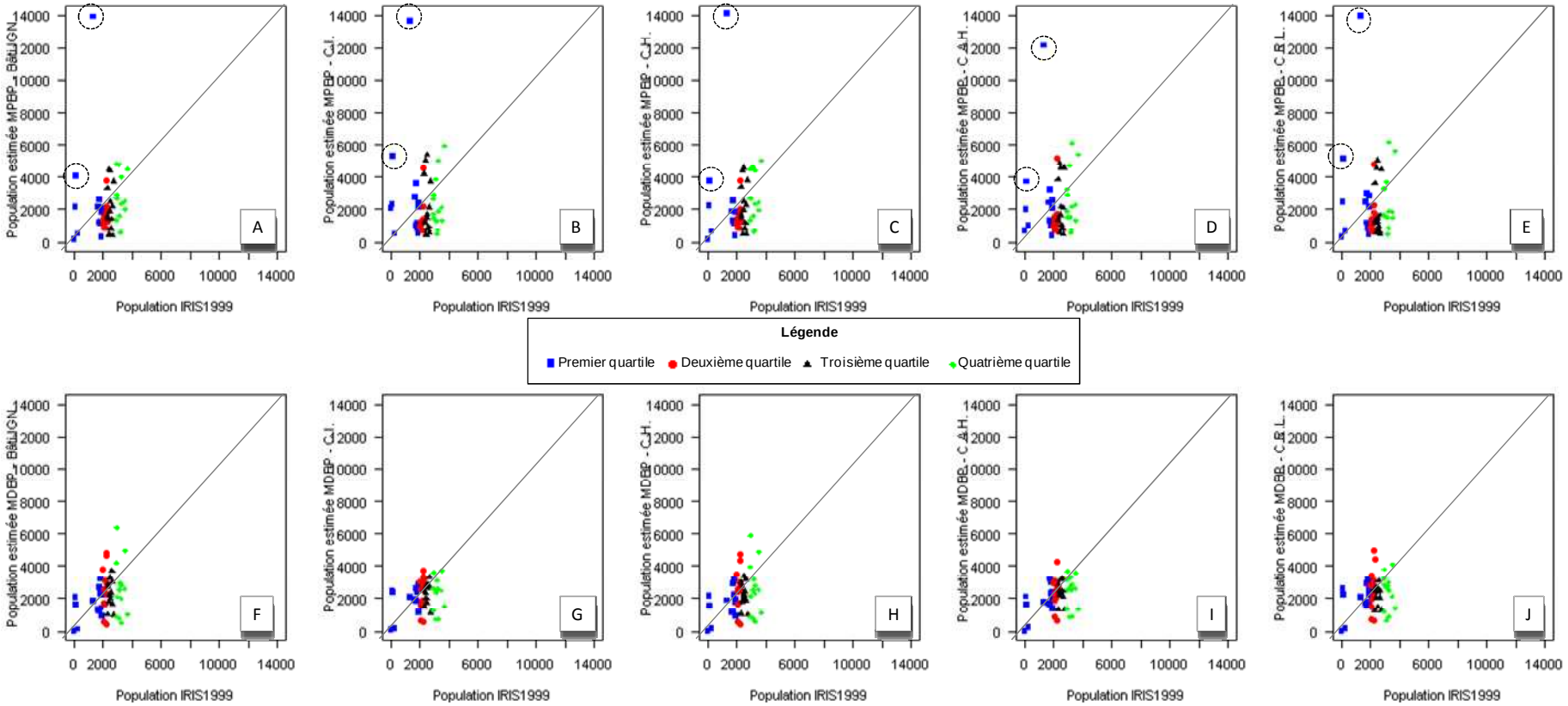
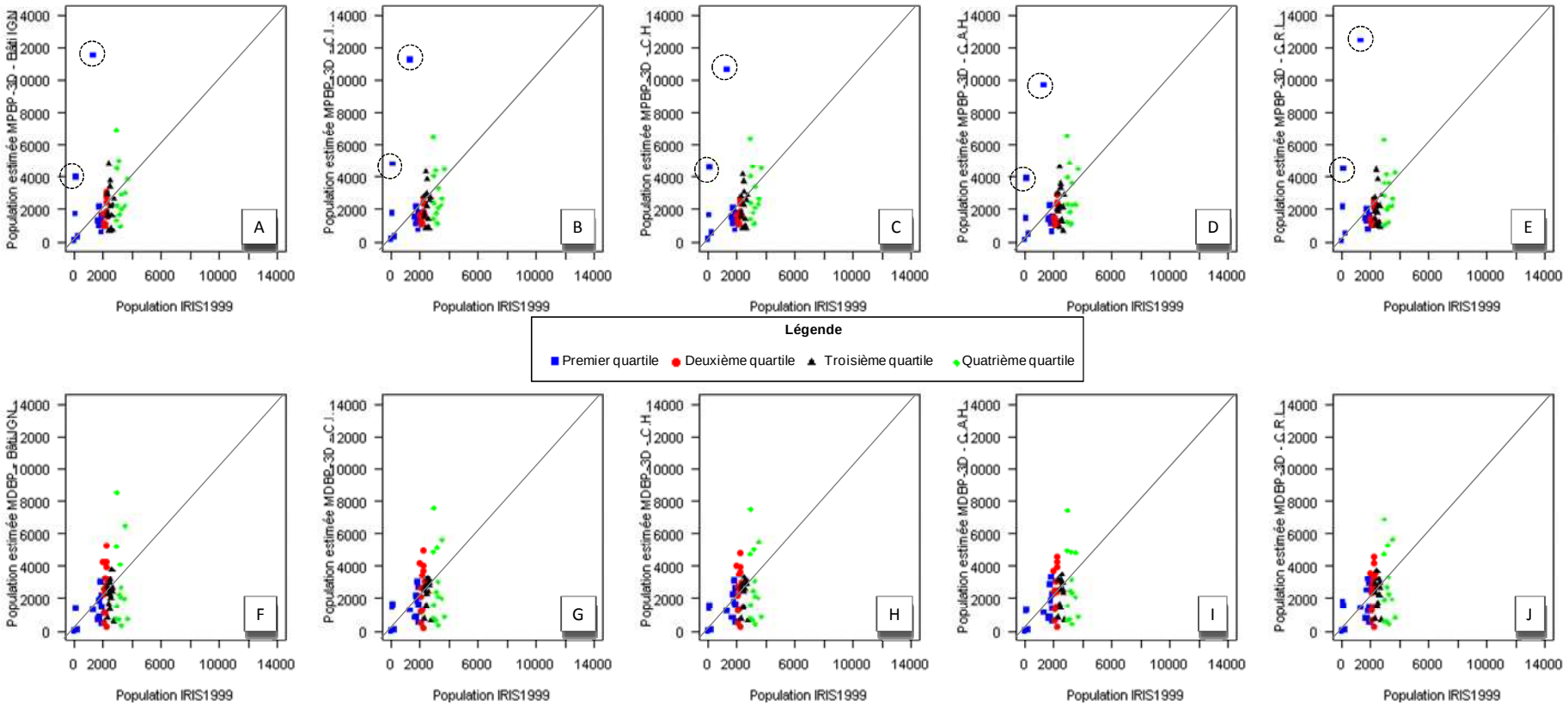


Figure 4-11 : Diagrammes bi-variés des populations estimées par l'approche volume *vs* la population des IRIS1999. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D



En résumé, on peut conclure que MPBP et MPBP-3D offrent les meilleurs résultats pour les IRIS (comme pour les îlots, voir annexe 4). Néanmoins, ces résultats ne sont pas complètement satisfaisants, car les valeurs des CCIs sont faibles, en particulier pour les IRIS.

4.3.2 Analyse de sensibilité

En raison des limites mises en évidence[50], une analyse de sensibilité a consisté à exclure les IRIS identifiés comme porteurs du biais dans la modélisation, en particulier ceux qui surestiment la population par une utilisation non résidentielle des sols.

Les résultats des CCIs de la nouvelle modélisation -excluant Châteaufarine et les Tilleroyes- sont présentés dans le tableau 4-2.

Tableau 4-2 : Coefficient de corrélation intra-classe pour les IRIS : N=50

Méthode de modélisation de la population		**Méthode de désagrégation**			
		Pixel-bâti pondéré CCI (IC95%)	**p**	**Densité du bâti pondérée** CCI (IC95%)	**p**
Approche par surface	Bâti IGN	0,32 (0,05-0,55)	0,01	0,24 (-0,03-0,49)	0,04
	C. ISODATA	0,16 (-0,12-0,42)	0,12	0,28 (0,00-0,52)	0,02
	C. Hiérarchisée	**0,33** (0,05-0,55)	0,01	0,27 (-0,00-0,51)	0,03
	C. Ascendante Hiérarchique	0,23 (-0,04-0,48)	0,05	0,33 (0,06-0,56)	0,008
	C. Régression logistique	0,22 (-0,05-0,47)	0,05	0,25 (-0,03-0,49)	0,05
Approche par volume	Bâti IGN	0,35 (0,09-0,57)	0,005	0,23 (-0,05-0,47)	0,03
	C. ISODATA	0,41 (0,15-0,61)	0,001	0,25 (-0,03-0,49)	0,04
	C. Hiérarchisée	**0,42** (0,16-0,62)	0,001	0,26 (-0,02-0,50)	0,03
	C. Ascendante Hiérarchique	0,40 (0,14-0,61)	0,001	0,26 (-0,02-0,50)	0,03
	C. Régression logistique	0,40 (0,14-0,61)	0,001	0,26 (-0,02-0,50)	0,03

[50] Les mêmes limites ont été mises en évidence pour les îlots, et donc les mêmes analyses de sensibilité ont été effectuées. Plus de détails sont donnés dans l'annexe 4.

Plusieurs enseignements sont à tirer de l'analyse de ces résultats.

D'une part, les valeurs du CCI n'ont quasiment pas changé pour MDBP ni par l'approche « surface », ni par l'approche « volume ». D'ailleurs, ces résultats sont plus faibles que ceux obtenus avec la MPBP.

D'autre part, les valeurs du CCI, obtenus par MPBP, ont significativement augmenté tant dans l'approche « surface » -passant d'un intervalle de [-0,05-0,03] à un intervalle de [0,16-0,32]- que dans l'approche « volume » -avec une augmentation de [0,05-0,13] à [0,35-0,42].

De plus, la quasi-totalité des estimations sont statistiquement significatives, hormis celle réalisée par l'approche « surface » avec la MPBP utilisant le « bâti » extrait par classification ISODATA. D'ailleurs, cette modélisation donne les valeurs les plus faibles du CCI (à savoir 0,16).

En outre, le « bâti » extrait par la classification hiérarchisée (§3.1.2) atteint les valeurs les plus élevées d'CCI, surpassant celles obtenues par le bâti d'IGN (donnée de référence) ; et ce dans l'approche « surface » (0,33 *vs* 0,32) comme dans l'approche « volume » (0,42 *vs* 0,35).

Pour finir, les résultats obtenus avec les « bâtis » extraits par la stratégie orientée-objets, -soit non dirigée (C.A.H.), soit dirigée (R.L)- sont semblables dans MPBP et dans MPBP-3D. Ces résultats sont, d'ailleurs, plus performants que ceux obtenus par classification ISODATA, mais moins satisfaisants que ceux obtenus par classification hiérarchisée.

Une deuxième analyse de sensibilité a été réalisée avec la méthode de désagrégation du pixel-bâti pondérée, apparue comme la plus performante dans l'analyse précédente. Inspirée des travaux de Li et Weng (2005), notre zone d'étude (excluant les IRIS écartés dans l'analyse précédente) a été divisée en trois groupes, en fonction de la densité de la population par unité d'analyse (fig. 4-1), à savoir : densité basse, densité moyenne et densité haute (tab. 4-3). Puis, la population a été recalculée à l'intérieur de chaque groupe, ainsi que la valeur du coefficient de corrélation intra-classe. Ces résultats apparaissent dans le tableau 4-4 pour les IRIS.

Tableau 4-3 : Distribution des unités géographiques en fonction de la densité de la population

Unité géographique	**Densité basse** Habitantes /Nb. unité	**Densité moyenne** Habitantes/Nb. unité	**Densité haute** Habitantes /Nb. unité
IRIS	32 768 hab./16 IRIS	40 809 hab./17 IRIS	42 643 hab./17 IRIS

Tableau 4-4 : Coefficient de corrélation intra-classe pour les IRIS en fonction de la densité de la population

		Méthode de désagrégation du pixel-bâti pondérée		
Méthode de modélisation de la population		**Densité basse** CCI (IC95%) N= 16	**Densité moyenne** CCI (IC95%) N= 17	**Densité haute** CCI (IC95%) N= 17
Approche par surface	Bâti IGN	0,73 (0,38-0,90)	0,53 (0,08-0,80)	0,12 (-0,36-0,55)
	C. ISODATA	0,65 (0,24-0,86)	0,66 (0,28-0,86)	0,16 (-0,33-0,59)
	C. Hiérarchisée	0,74 (0,40-0,90)	0,57 (0,13-0,81)	0,15 (-0,34-0,57)
	C. Ascendante Hiérarchique	0,77 (0,45-0,91)	0,70 (0,34-0,88)	0,19 (-0,31-0,61)
	C. Régression logistique	0,74 (0,39-0,90)	0,50 (0,04-0,78)	0,14 (-0,35-0,57)
Approche par volume	Bâti IGN	0,68 (0,29-0,87)	0,27 (-0,23-0,65)	0,17 (-0,32-0,59)
	C. ISODATA	0,74 (0,40-0,90)	0,32 (-0,17-0,68)	0,20 (-0,29-0,61)
	C. Hiérarchisée	0,76 (0,43-0,91)	0,34 (-0,16-0,69)	0,21 (-0,28-0,62)
	C. Ascendante Hiérarchique	0,76 (0,43-0,91)	0,32 (-0,17-0,69)	0,20 (-0,29-0,62)
	C. Régression logistique	0,71 (0,34-0,89)	0,34 (-0,15-0,70)	0,26 (-0,24-0,65)

Après analyse de ces résultats, nous pouvons déduire que, de manière globale, les valeurs de corrélations sont plus fortes (sauf pour la densité haute) quand la modélisation s'effectue sur des surfaces avec une densité de population plus homogène. Néanmoins, la modulation morphologique de la densité[51], c'est-à-dire les différentes formes urbaines pour

[51] Par exemple, on peut retrouver la même densité du bâti avec différentes combinaisons entre l'emprise au sol et la hauteur : faible emprise au sol et grande hauteur ; emprise au sol moyenne et hauteur moyenne ; ou

la même densité bâtie, joue un rôle important dans la modélisation. Ainsi, l'effet de la hauteur est moindre quand les zones à modéliser ont une densité du bâti homogène, et inversement, il se révèle plus important quand ce n'est pas le cas. Cela explique la faible corrélation obtenue avec la catégorie « densité haute » : cette dernière catégorie est constituée de zones ayant de grandes hauteurs -à Planoise- mais aussi des zones ayant de fortes emprises au sol -au centre ville.

Pour conclure, la modélisation de la population effectuée par méthode dasymétrique, utilisant comme donnée auxiliaire la variable « bâti/non bâti » (construite à partir de la classification des données télédétection THRS) permet de mettre en évidence différents points.

En premier lieu, l'indicateur de répartition le plus simple, le « pixel-bâti », c'est-à-dire la somme des pixels par unité géographique (voir 4.1.1) se révèle le plus performant.

En deuxième lieu, l'introduction de la hauteur (c'est-à-dire l'approche « volume »), améliore la performance de la méthode dasymétrique, notamment, dans des zones ayant différentes formes urbaines pour une même densité du bâti. Ainsi, nos résultats rejoignent ceux des dernières études qui ont employé une approche volumétrique (Dong. *et al.*, 2010 ; Lu *et al.*, 2010; Lwin et Murayama, 2011; Wu *et al*. 2008).

L'analyse de sensibilité démontre que cibler la zone d'étude à modéliser (c'est-à-dire identifier au préalable la zone résidentielle, ou à l'inverse ne pas prendre en compte les zones non résidentielles) permet d'obtenir de meilleurs résultats. En effet, la méthode de « la variable limitante » a été rapportée comme la méthode la plus performante (Eicher et Brewer, 2001 ; Langford *et al.,* 2008).

L'utilisation de deux échelles spatiales (IRIS et îlot, ce dernier étant détaillé dans l'annexe 4) a permis de constater la convergence des résultats, après les analyses de sensibilité (modélisation sans les zones commerciales, industrielles et institutionnels ; modélisation par densité de population). C'est-à-dire que les résultats les plus significatifs ont été obtenus avec les mêmes méthodes MPBP et MPBP-3D et avec les mêmes limites. L'agrégation des

encore, forte emprise au sol et faible hauteur (Institut d'aménagement et d'urbanisme de la région d'Ile de France, 2005).

données est possible avec l'emboîtement d'échelles successives vers les échelles supérieures (des îlots vers les IRIS) (Joly, 1990), même si demeure le problème de l'unité aréolaire modifiable « MAUP » (*Modifiable Areal Unit Problem*, en anglais) (Openshaw, 1984). Il s'ensuit que le niveau de détail que nous pourrons atteindre avec la méthode dasymétrique est fortement lié au niveau de détail des données auxiliaires utilisées pour désagréger la population. Cela explique que, dans la recherche menée par Viel et Tran (2009), les résultats obtenus ont été plus adaptés au niveau des IRIS qu'au niveau des îlots. En effet, le niveau de détail fourni par une image Landsat (30m de résolution spatiale) n'est pas suffisant pour une analyse à l'échelle des Îlots : il se produit de la « saturation » dans les zones à habitat dispersé (*urban sprawling*, en anglais). En outre, les résultats obtenus, dans notre modélisation, avec ces deux niveaux d'analyse, n'ont pas été tout à fait identiques. Et ce parce que : d'une part, le niveau de détail de l'échelle ne permet pas la même délimitation des zones non résidentielles (avec les îlots, nous avons écarté du calcul une fraction de l'IRIS, tandis qu'avec les IRIS nous avons exclu la totalité de l'unité géographique) ; d'autre part, la modulation morphologique de la densité est différente. Néanmoins, nous avons établi que la méthode dasymétrique s'adapte facilement aux différentes contraintes tout en procurant de bons résultats. Cela garantit d'une certaine manière la reproductibilité de la méthode dans des contextes différents, par exemple, la modification des aires à modéliser, ou encore le changement de l'échelle ciblée, répondant aux attentes des épidémiologistes.

Nous avons présenté dans ce chapitre deux approches différentes, « surface » et « volume », pour modéliser la population à partir de la variable « bâti / non bâti » issue des données THRS. En sus, chacune de ces deux approches a utilisé deux méthodes de répartition des données différentes, à savoir : le pixel bâti -fondé sur la somme des pixels-, et la densité du bâti -fondé sur la densité des pixels bâtis-; et ce en utilisant l'unité géographique des IRIS comme frontière de cette estimation de la population à échelle infra-communale. Au total, 20 cartes de la population estimée ont été produites, en combinant ces différentes approches avec la variable « bâti / non bâti », et elle aussi avec cinq versions différentes (en fonction de la méthode d'extraction des bâtiments). En outre, les résultats de ces estimations ont été validés avec la population de référence -le recensement de l'INSEE 1999. Pour ce faire, le CCI a été calculé, et des graphiques bi-variés (population estimée *versus* population observée) ont été réalisés. De plus, en raison des résultats obtenus, et afin d'établir la portée de la méthode, des analyses de sensibilité ont été conduites. *In fine*, ces résultats ont mis en évidence, entre autres, les bénéfices à inclure la hauteur dans la modélisation de la population, ainsi que la facilité d'adaptation de la méthode dasymétrique.

« On attribue habituellement à John Snow la première analyse épidémiologique spatiale au niveau local, à l'occasion d'une épidémie de choléra à Londres (1848) : la répartition préférentielle des décès dans la zone urbaine desservie par l'une des deux compagnies de distribution d'eau a permis d'identifier l'origine de cette épidémie » Czernichow et son équipe (Épidémiologie, 2001)

Chapitre 5. Estimation des taux bruts d'incidence : application de l'épidémiologie descriptive

La troisième étape de la méthodologie proposée concerne le calcul des taux bruts d'incidence, en utilisant au numérateur les populations estimées -par la méthode dasymétrique- dans l'étape précédente (cf. chapitre 4). Pour obtenir un nombre suffisant de cas au numérateur, seule l'échelle des IRIS est pertinente. Nous disposons des cas d'incidence de deux types de cancer différents (cf. 2.3.3) : d'une part, le cancer du sein (chez les femmes), qui est relativement fréquent ; et d'autre part, le lymphome non-hodgkinien, plutôt rare. Ainsi, chacune des vingt populations estimées pour les IRIS (cf. 4.2) sont utilisées dans le but de calculer les taux bruts d'incidence de ces deux maladies.

Par ailleurs, une brève présentation de chacune de ces deux maladies est faite au début de ce chapitre, afin de mieux comprendre l'importance de l'estimation des taux d'incidence en épidémiologie descriptive. Ensuite, les cartes des taux bruts d'incidence estimés sont présentées. Puis ces estimations sont validées à travers le calcul du coefficient de corrélation intra-classe (CCI), prenant en compte les taux d'incidence calculés avec le recensement comme données de référence. Ce chapitre finit sur trois différentes analyses de sensibilité de la méthode, qui ouvrent des perspectives dans l'utilisation de la télédétection pour des applications en santé publique.

5.1 Un aperçu de l'épidémiologie du cancer

« Le cancer est un ensemble de cellules anormales qui se multiplient de façon incontrôlée. Elles finissent souvent par former une masse qu'on appelle tumeur maligne » (FNCLCC[52], 2007). Selon l'Inserm (2008), les cancers figurent parmi les pathologies pouvant être liées à l'environnement dont les modifications pourraient être partiellement responsables de l'augmentation constatée de l'incidence de certains cancers. Selon la base de données GLOBOCAN 2008 (Ferlay *et al.*, 2010), 12 662 600 nouveaux cas de cancer et 7 564 800 décès ont été registrés au niveau mondial, en 2008. D'ailleurs, en France, en 2005, le cancer (toutes localisations confondues) se situe au 1er rang des décès chez les hommes et au 2ème rang des décès chez les femmes (Inserm, 2008). De plus, selon les travaux de Jemal et de son équipe (2011) –s'appuyant sur GLOBOCAN 2008-, les taux d'incidence, tous cancers confondus (toutes localisations, chez les hommes comme chez les femmes), sont quasiment deux fois plus élevés dans les pays économiquement développés que dans les pays en développement. Une telle disparité peut trouver son origine dans une hétérogénéité de la prévalence des principaux facteurs de risque et des pratiques de détection. Cela révèle l'importance de la quantification de l'incidence du cancer pour la mise en place d'une politique de prévention, et pour une meilleure évaluation des besoins en termes de prise en charge de la maladie (Tourancheau, 2008).

5.1.1 Le cancer du sein chez les femmes : une maladie plutôt fréquente

Le cancer du sein est une tumeur maligne qui touche la glande mammaire ; son étiologie reste pour une part inconnue, mais de nombreux facteurs de risque -génétique, hormonal et/ou environnemental et comportemental- ont été identifiés (Inserm, 2008 ; Woronoff et Danzon, 2008). Néanmoins, les avancées moléculaires et cellulaires devraient faire progresser la connaissance de l'oncogenèse mammaire (conversion d'une cellule normale en cellule tumorale - Inserm, 2008).

[52] Fédération Nationale des Centres de Lutte Contre le Cancer.

Au niveau mondial, le cancer du sein a touché 1 643 000 nouveaux cas en 2010, contre 641 000 nouveaux cas en 1980, avec un taux d'accroissement annuel de 3,1% (Forouzanfar *et al.*, 2011). Toutefois, cet accroissement observé au cours des dernières décennies est, schématiquement, en grande partie attribué au développement du dépistage dans les pays « industrialisés » (Inserm, 2008). En 2010, le nombre de décès dus à ce cancer a été, dans les pays « développés », de 425 000 femmes, dont 68 000 entre 15 et 49 ans (Forouzanfar *et al.*, 2011).

Selon l'Observatoire de la Santé en Franche Comté et le registre des tumeurs du Doubs, en 2005 dans la région (Woronoff et Danzon, 2008), le cancer le plus fréquent chez les femmes était le cancer du sein, avec 874 nouveaux cas. De plus, entre 1980 et 2005, le nombre annuel de nouveaux cas de cancer du sein est passé de 384 à 874, c'est-à-dire une augmentation du 77% des taux standardisés d'incidence[53]. L'évolution sur la totalité du territoire français est similaire, l'augmentation atteignant 79%. Le taux standardisé d'incidence au niveau régional est de 125,7 pour 100 000 femmes, le taux national atteignant 134,5. Quant au taux de décès, en 2005, ce cancer représentait 19% des décès avec 200 cas par année. Néanmoins, dans la période 1981-2004, une baisse de 12% des taux standardisés de mortalité a été registrée, tandis qu'au niveau national cette diminution atteint seulement 4%. Ainsi, le taux de décès en Franche-Comté est 28,0 pour 100 000 femmes ; et au niveau national de 29,3 pour 100 000 femmes. En conclusion, l'incidence du cancer du sein a augmenté de façon importante et constante depuis 25 ans alors que la mortalité a tendance à diminuer, se situant à un niveau bien inférieur à celui de l'incidence (Inserm, 2008 ; Woronoff et Danzon, 2008). Ce contraste s'explique, en partie, par l'amélioration des thérapeutiques et par un diagnostic plus précoce (Inserm, 2008).

Le cancer du sein se situe donc au 1er rang des cancers de la femme aux niveaux : mondial (Forouzanfar *et al.*, 2011), européen (Inserm, 2005), national (Inserm, 2008), ainsi que départemental (Woronoff et Danzon, 2008).

[53] « Le taux standardisé permet de comparer des groupes qui diffèrent par leur milieu et leur structure notamment pour l'âge ». « La standardisation vise à tenir compte des effectifs des différents groupes composant une population pour pouvoir comparer des taux entre eux. La méthode directe consiste à appliquer, à une population de référence, les taux spécifiques de la population étudiée. La méthode indirecte consiste à appliquer les taux spécifiques par groupe ou classe d'une population de référence aux effectifs des mêmes groupes de la population étudiée. » (consulté le 11 juin 2012)
http://acces.ens-lyon.fr/acces/ressources/sante/epidemiologie/GlossairEpidem/GlossEpidThemes

5.1.2 Le lymphome non-hodgkinien : un cancer assez rare

Les lymphomes sont des cancers du système lymphatique[54], lequel participe aux réactions de défense de l'organisme. L'immunodépression congénitale ou acquise est un facteur étiologique reconnu ; le rôle de certains virus (de l'hépatite, Epstein-Barr) est également à prendre en considération (Woronoff et Danzon, 2008). Mais la grande majorité des lymphomes reste d'origine inexpliquée. Selon la classification de l'OMS (2008), il existe plusieurs types de lymphomes non-hodgkiniens (LNH), qui présentent des évolutions différentes (Inserm, 2005). En outre, le LNH s'observe à des taux faibles dès 5 ans, puis sa fréquence augmente avec l'âge (Inserm, 2005 ; Woronoff et Danzon, 2008).

Au niveau mondial, en 2008, on estime que le LNH a touché 355 900 nouveaux cas, à savoir 199 600 hommes et 156 300 femmes, se situant ainsi au 8ème rang des cancers chez les hommes et au 10ème rang des cancers chez les femmes (Jemal *et al.*, 2011). Le taux d'incidence du lymphome non hodgkinien a augmenté dans la plupart des pays développés au cours des années 1990, se stabilisant ces dernières années. Selon Jemal et son équipe (2011), les augmentations avant 1990 peuvent être attribuées en partie à l'amélioration des procédures de diagnostic et des changements dans la classification, ainsi qu'à l'apparition du Syndrome d'Immunodéficience Acquise (SIDA). En ce qui concerne le nombre de décès, en 2008, ce cancer a été la cause de 191 400 morts, dont 109 500 chez les hommes.

Au niveau européen, selon l'étude menée par Institut national de la santé et de la recherche médicale (Inserm) en 2005, le taux d'incidence du lymphome malin non-hodgkinien (LMNH[55]) standardisé sur la population mondiale est de 13,3/100 000 chez l'homme et de 7,8/100 000 chez la femme. Ce cancer se situe au 7ème rang des décès par cancer et représente 3,5 % de l'ensemble des décès par cancer. Les taux de mortalité standardisés sont estimés à 5,3/100 000 chez l'homme et 3,4/100 000 chez la femme.

Selon l'Observatoire de la Santé en Franche Comté et le registre des tumeurs du Doubs, (Woronoff et Danzon, 2008) en Franche-Comté en 2005, le LMNH occupait le sixième rang des cancers, touchant une centaine de nouveaux cas chez les hommes (101 nouveaux cas),

[54] « Se dit du réseau de petits vaisseaux et de ganglions qui transportent la lymphe ». « La lymphe est un liquide légèrement coloré produit par le corps dans lequel baignent les cellules. La lymphe transporte et évacue les déchets des cellules ». (FNCLCC, 2007)

[55] Voir « lymphome non-hodgkinien »

et un chiffre proche chez les femmes (86 nouveaux cas). Cela représentait 3% des cancers chez l'homme et 4% chez la femme. Le taux standardisé d'incidence régional atteignait chez les hommes 15,3 pour 100 000, contre 16,1 pour 100 000 pour toute la France. En revanche, chez les femmes, ce taux régional se situait à 10,9 pour 100 000, et pour toute la France 11 pour 100 000. L'augmentation annuelle moyenne des nouveaux cas de ce cancer entre 1978 et 2004 correspond à 3,1% chez les hommes et à 3,9% chez les femmes. Néanmoins, une décélération s'est produite chez les hommes, en particulier sur la période 2000-2005. Quant aux décès, le LMNH se situait au 12ème rang des décès des cancers chez les hommes (taux standardisés), tandis qu'il se situait au 7ème chez les femmes. Entre 1984 et 2005, les taux standardisés de mortalité par LMNH ont fortement progressé, chez les hommes comme chez les femmes.

En conclusion, l'incidence et la mortalité ont augmenté de façon marquée au cours des deux dernières décennies, tant au niveau européen (Inserm, 2005), qu'au niveau départemental (Woronoff et Danzon, 2008). En revanche, l'étude au niveau global (Jemal *et al.*, 2011) met en avance, d'un côté, une incidence haute dans les pays développés ; de l'autre, une incidence basse dans les pays en développement.

5.2 Estimation des taux bruts d'incidence

Nous nous appuyons sur les définitions « d'incidence » et de « taux d'incidence », proposées par Bouyer et son équipe (2005), afin d'introduire ce sous-chapitre.

> *Les mesures d'incidence quantifient la « production » de nouveaux cas de maladie dans la population. …Pour que l'incidence ait un sens, il faut tout d'abord préciser sur quelle période de temps la production de nouveaux cas est enregistrée. En second lieu, il ne faut pas oublier que seuls les non-malades sont susceptibles de « produire » des nouveaux cas.*
> *Par définition, le taux d'incidence TI de la maladie est la « vitesse de production » de nouveaux cas. Il est donc égal au nombre de nouveaux cas par unité de temps divisé par la taille de la population.*

Ainsi le taux d'incidence est défini par :

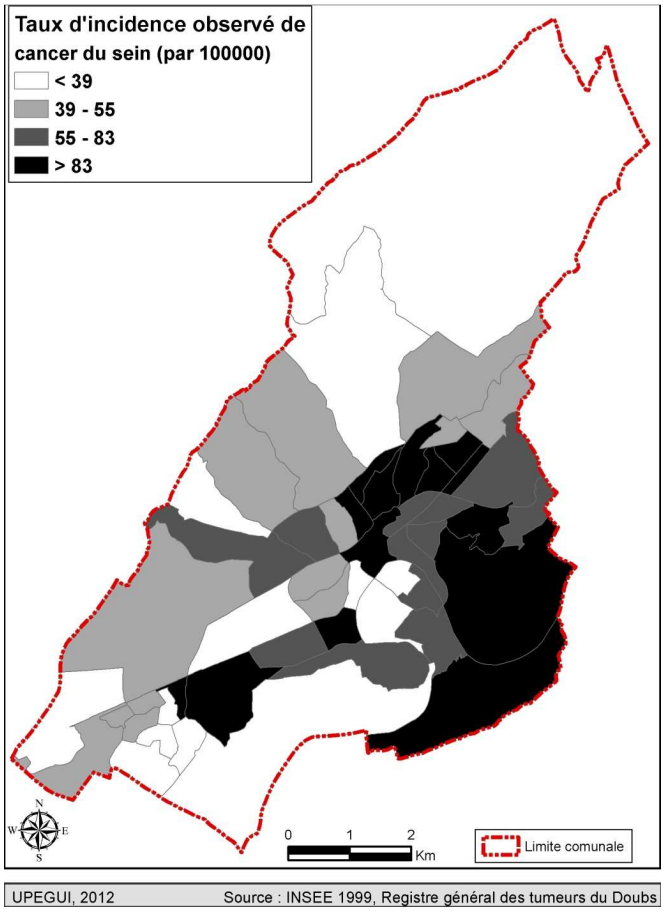
$$Taux\ d'Incidence = \frac{nombre\ de\ nouveaux\ cas\ pendant\ la\ periode\ donnée}{taille\ de\ la\ population}$$

Certes, différentes unités de mesures peuvent être appliquées à la taille de la population (personnes-temps), néanmoins avec la méthodologie proposée, la taille de la population est limitée au nombre de sujets. Nous nous trouvons donc dans le cadre d'une population ouverte, c'est-à-dire en constant renouvellement (naissances et emménagements *vs* décès et déménagements) où « il est fréquent que ses caractéristiques moyennes (d'âge notamment) restent stables au cours du temps » (Bouyer *et al.*, 2005). Nous estimons ainsi seulement des taux bruts d'incidence, ce qui signifie, qu'on calcule le rapport (nouveaux cas / population) sans considération d'autres facteurs : ni la composition de la population, ni la cause de l'événement (Bernard et Lapointe, 1987).

5.2.1 Le cancer du sein chez les femmes

Dans la pratique, les incidents de cancer du sein (491 cas - chez les femmes) diagnostiqués à Besançon entre 1996 et 2002 (Viel *et al.*, 2008) ont été cumulés. Ensuite, les taux d'incidence observés ont été calculés (fig. 5-1) en utilisant ces cas incidents -comme numérateur- et les chiffres du recensement de 1999 -comme dénominateur.

Figure 5-1 : Taux d'incidence observés dans les cancers du sein, période entre 1996 et 2002



Ces taux constituent la référence pour valider les résultats obtenus lors des calculs des taux estimés, en utilisant comme dénominateur les populations obtenues lors de la modélisation par les différentes approches (cf. chapitre 4).

5.2.1.1 Estimation des taux à partir des populations issues de l'approche surface

Dans la figure 5-2, on trouve les taux bruts d'incidence calculés utilisant les différentes populations estimées : d'une part, par la méthode du pixel bâti-pondérée (de A à E) ; et d'autre part, par la méthode de la densité du bâti pondérée (de F à J).

Dans les deux méthodes, la discrétisation des taux est effectuée avec les quartiles des taux de référence, conduisant aux intervalles : c'est-à-dire < 39 ; 39-55 ; 55-83 ; > 83 (fig. 5-1).

Pour les cartes, nous observons, globalement, une tendance différente entre les deux méthodes. Aussi, les taux bruts estimés à partir de la méthode du pixel bâti pondérée (MPBP) présentent-ils les valeurs les plus élevées, concentrées surtout dans deux parties de la ville : d'un côté, sur les IRIS localisés à Planoise ; et de l'autre, sur les IRIS localisés au nord du centre-ville, entre les quartiers des Chaprais et la deuxième couronne sur la plaine. Pour les taux estimés avec la méthode de la densité du bâti pondérée (MDBP), la distribution des valeurs les plus élevées change. Celles-ci se concentrent ainsi vers l'extérieur de la ville : tant au nord, sur les zones à habitat dispersé (sauf à Chailluz) ; qu'au sud, dans la deuxième couronne sur la colline.

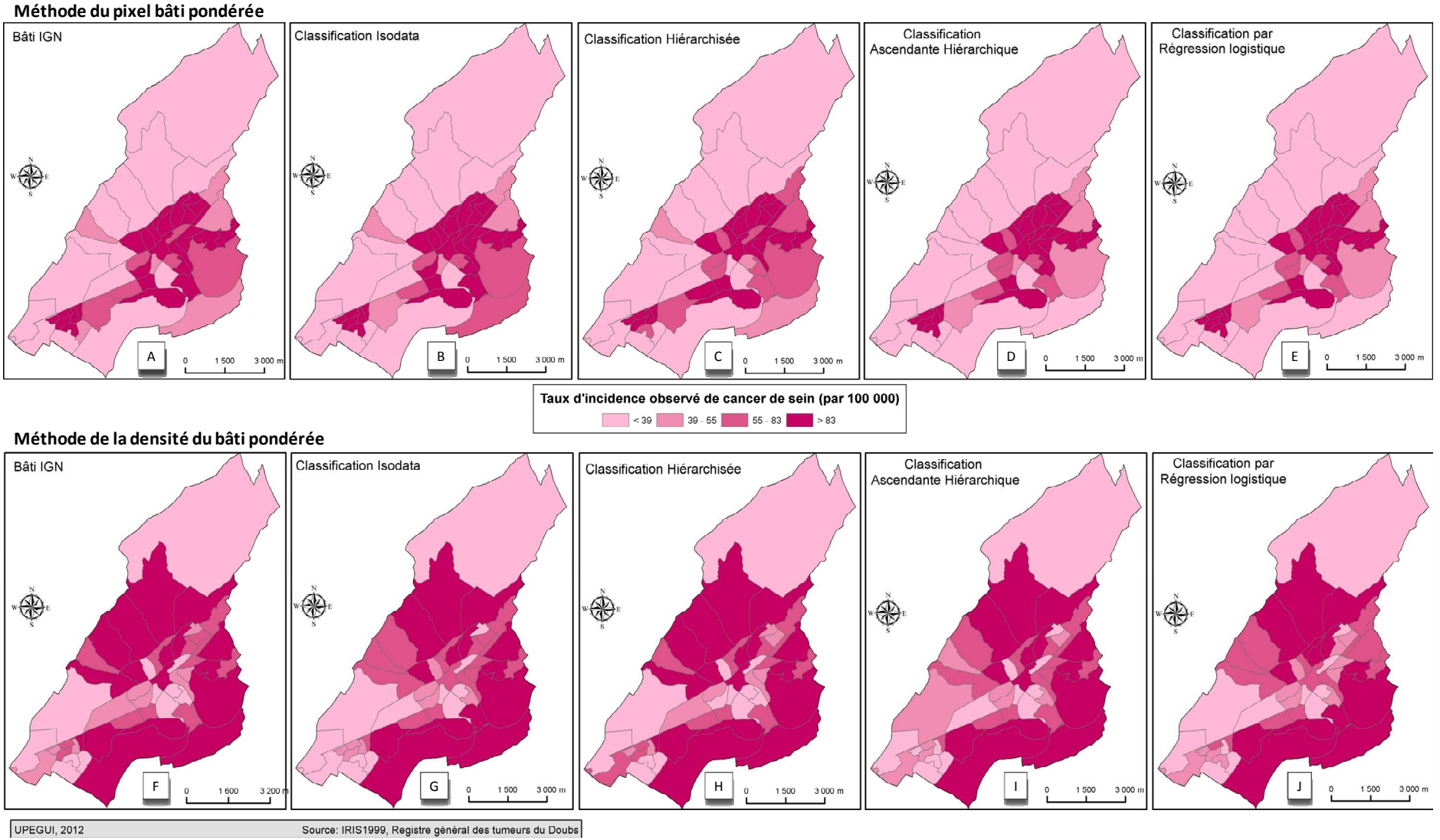
Néanmoins, en analysant plus en détail, nous observons des variations dans la distribution des IRIS à l'intérieur de chacune des méthodes (tab. 5-1). Pour la MPBP, le premier intervalle (< 39) a un nombre d'IRIS assez homogène dans les cinq types de bâti. Le deuxième intervalle (entre 39 et 55) concerne le nombre d'IRIS le plus faible, en particulier pour les populations estimées à partir du bâti de l'IGN (fig. 5-2A) et celui extrait par la classification hiérarchisée (fig. 5-2C). Le troisième intervalle (entre 55 et 83) est le plus hétérogène dans la distribution du nombre d'IRIS, avec d'un côté, un nombre plus important pour les populations estimées à partir du bâti de l'IGN (fig. 5-2A) et de la classification hiérarchisée (fig. 5-2C) ; et de l'autre côté, un nombre plus faible pour les populations estimées à partir du bâti de la classification ISODATA (fig. 5-2B), ascendante hiérarchisée (fig. 5-2C) ainsi que par régression logistique (fig. 5-2E). Le dernier intervalle (> 83) est plutôt

homogène et comprend un nombre important d'IRIS. Pour la MDBP, la distribution du nombre des IRIS dans chacun des intervalles est plutôt homogène. Toutefois, les nombres les plus faibles se rencontrent dans le deuxième intervalle (entre 39 et 55), en particulier avec la population estimée à partir du bâti provenant de la classification hiérarchisée (fig. 5-2H) ; cependant au contraire le nombre le plus élevé dans cet intervalle correspond à la population estimée à partir du bâti issu de la classification par régression logistique (fig. 5-2J). En revanche, les effectifs les plus importants se trouvent dans le troisième (entre 55 et 83) et dans le dernier intervalle (> 83).

Tableau 5-1 : Fréquence des IRIS par méthode d'estimation de la population et par intervalle du taux de cancer du sein

		Population issue de la méthode du pixel-bâti pondérée			
Intervalle du taux d'incidence de cancer du sein		**< 39**	**39 - 55**	**55 – 83**	**> 83**
Source du bâtiment	Bâti IGN	15	3	12	22
	Classification ISODATA	14	5	5	28
	Classification Hiérarchisée	15	3	11	23
	Classification Ascendante Hiérarchique	16	5	6	25
	Classification Régression logistique	13	6	6	27
		Population issue de la méthode de la densité du bâti pondérée			
Intervalle du taux d'incidence de cancer du sein		**< 39**	**39 - 55**	**55 – 83**	**> 83**
Source du bâtiment	Bâti IGN	12	9	14	17
	Classification ISODATA	13	8	17	14
	Classification Hiérarchisée	13	8	14	17
	Classification Ascendante Hiérarchique	14	9	15	14
	Classification Régression logistique	12	12	14	14

Figure 5-2 : Taux d'incidence estimés des cancers du sein, fondés sur la population modélisée par l'approche surface. De A à E par la méthode MPBP ; de F à J par la méthode MDBP



5.2.1.2 Estimation des taux à partir des populations issues de l'approche volume

Les taux bruts d'incidence calculés en utilisant les différentes populations estimées par l'approche volume se trouvent dans la figure 3. Ainsi, les figures de 3A à 3E correspondent aux taux évalués à partir des populations estimées par la méthode du pixel bâti-pondérée ; tandis que les figures de 3F à 3J représentent ceux calculés à partir des populations estimées par la méthode de la densité du bâti pondérée. Par ailleurs, la légende de ces cartes des taux estimés est la même que celle du taux de référence et que celle des cartes issues de l'approche surface.

Nous observons sur les cartes, de manière globale, une tendance différente pour chacune de deux méthodes ; cette tendance étant d'ailleurs similaire à celle de l'approche surface. Pour les taux bruts estimés à partir de la méthode du pixel bâti pondérée en 3D (MPBP-3D) les valeurs les plus élevées de l'incidence se concentrent quasiment toutes sur deux parties de la ville : d'un part, sur les IRIS localisés au nord-est du centre-ville, notamment sur la deuxième couronne sur la plaine ; et d'autre part, sur les IRIS localisés au sud-ouest du centre-ville, en particulier sur la frontière entre la deuxième couronne sur la plaine et celle sur la colline, jusqu'à atteindre Planoise. Pour les taux bruts estimés avec la méthode de la densité du bâti pondérée en 3D (MDBP-3D), les valeurs d'incidence les plus élevées se trouvent quasiment toutes sur les IRIS localisés sur les deuxième et troisième couronnes de la ville à l'exception de Chailluz et Planoise.

En analysant plus en détail les différences dans la distribution des taux de cancers à l'intérieur de chaque méthode (tab. 5-2) nous remarquons plusieurs éléments.

Pour la MPBP-3D, le premier intervalle (< 39) possède un nombre d'IRIS plutôt homogène dans les cinq types de bâti. Dans le deuxième intervalle (entre 39 et 55) se trouvent les nombres d'IRIS les plus faibles, notamment pour le taux calculé en utilisant la population issue du bâti identifié par la classification par régression logistique (fig. 5-3E). En revanche, dans le troisième intervalle (entre 55 et 83) l'effectif le plus élevé correspond à la classification par régression logistique (fig. 5-3E) ; les autres effectifs se présentent d'une façon relativement homogène. Dans le dernier intervalle (> 83) se trouvent les nombres d'IRIS les plus élevés, ceux-ci étant de plus homogènes.

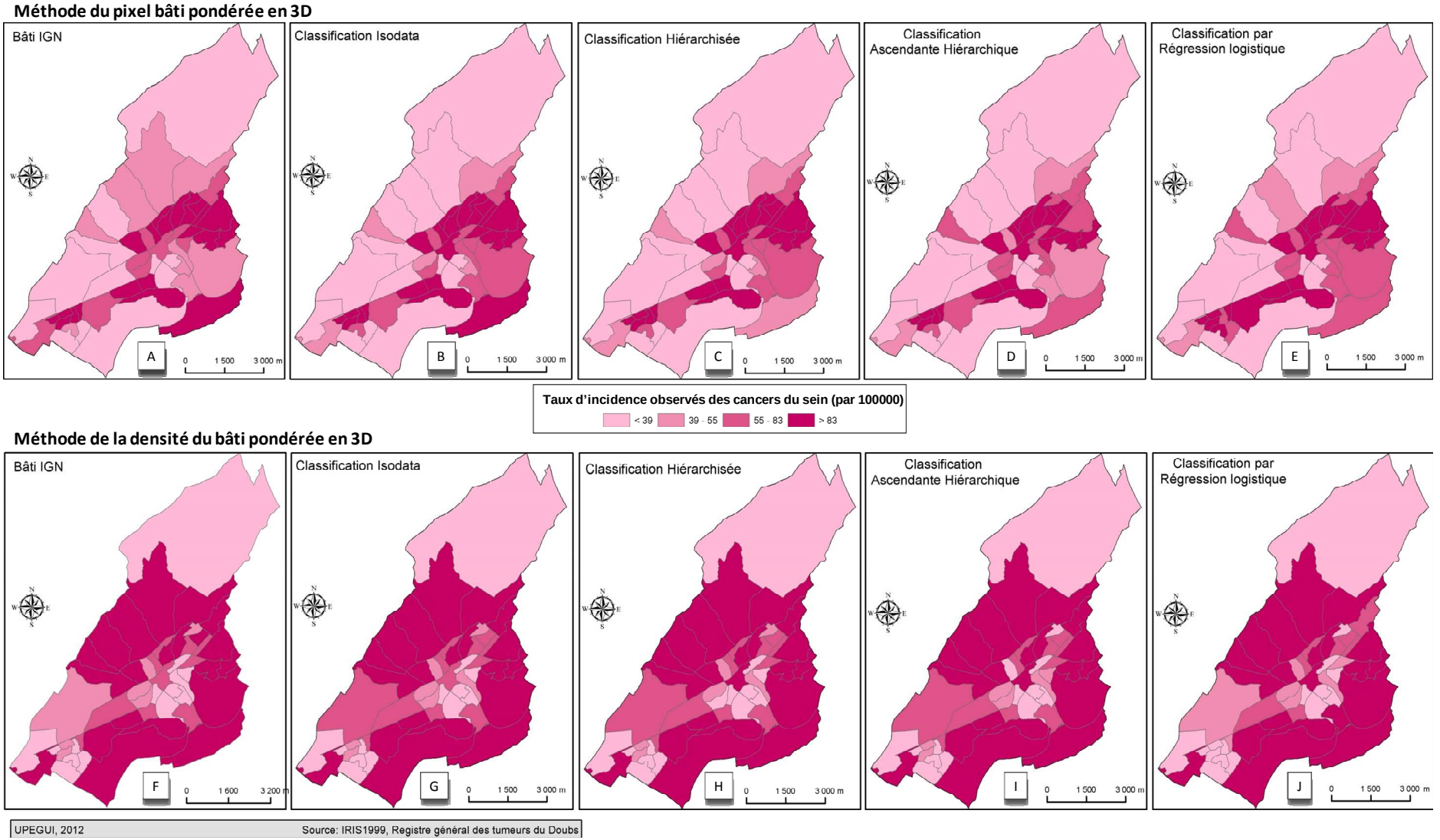
Pour la MDBP-3D, les effectifs des IRIS se distribuent plutôt de manière homogène dans chacun des intervalles. Toutefois, la classification par régression logistique conduit à un nombre important d'IRIS dans le premier intervalle aux dépens du troisième intervalle. Les effectifs les plus faibles se localisent dans le deuxième intervalle (entre 39 et 55), (phénomène qu'on constate d'ailleurs pour l'équivalent de cette méthode : avec l'approche surface), tandis que les effectifs les plus importants se trouvent dans le dernier intervalle (> 83).

Tableau 5-2 : Fréquence des IRIS par méthode d'estimation de la population et par intervalle du taux de cancers du sein

		Population issue de la méthode du pixel-bâti pondérée en 3D			
Intervalle du taux d'incidence des cancers du sein		**< 39**	**39 - 55**	**55 – 83**	**> 83**
Source du bâtiment	Bâti IGN	12	9	9	22
	Classification ISODATA	14	5	10	23
	Classification Hiérarchisée	14	6	10	22
	Classification Ascendante Hiérarchique	14	5	12	21
	Classification Régression logistique	12	3	15	22

		Population issue de la méthode de la densité du bâti pondérée en 3D			
Intervalle du taux d'incidence des cancers du sein		**< 39**	**39 - 55**	**55 – 83**	**> 83**
Source du bâtiment	Bâti IGN	14	7	10	21
	Classification ISODATA	13	7	12	20
	Classification Hiérarchisée	14	6	11	21
	Classification Ascendante Hiérarchique	14	6	11	21
	Classification Régression logistique	16	7	8	21

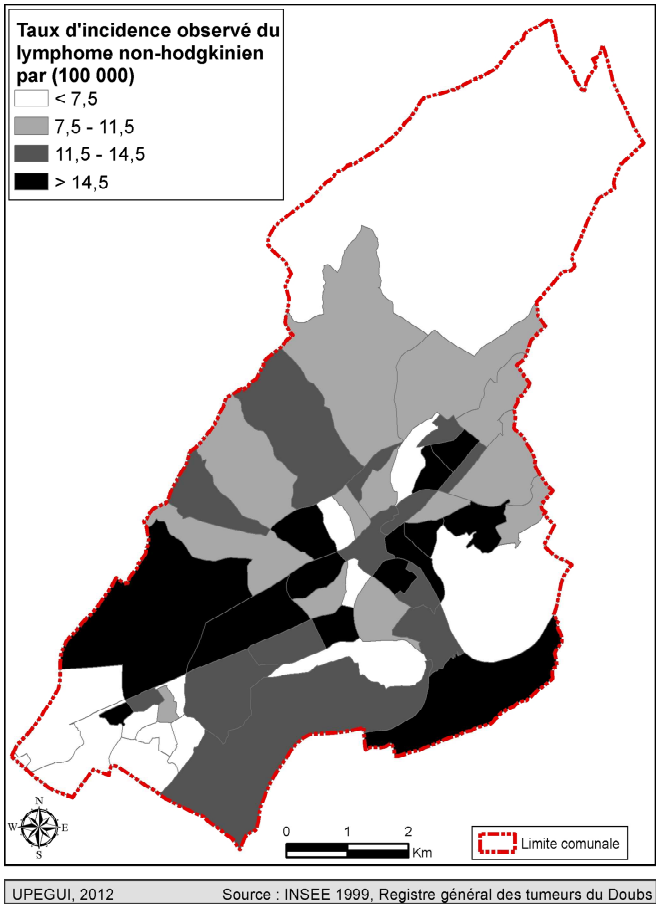
Figure 5-3 : Taux d'incidence estimés de cancers du sein, fondés sur la population modélisée par l'approche volume. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D



5.2.2 Le lymphome non-hodgkinien

Les 222 nouveaux cas de lymphomes non hodgkiniens inclus dans notre étude ont été diagnostiqués à Besançon pendant 16 ans, entre 1980 et 1995 (Floret *et al.*, 2003). Les taux d'incidence observé du LNH ont été calculés (fig. 5-4) avec le recensement de 1999 -comme dénominateur.

Figure 5-4 : Taux d'incidence observés du lymphome non-hodgkinien, période comprise entre 1980 et 1995



Sur la carte des taux d'incidence observés du LNH (fig. 5-4), on ne distingue pas de concentrations particulières, c'est-à-dire que la distribution de ce cancer est plutôt aléatoire sur toute l'étendue de la ville.

5.2.2.1 Estimation des taux à partir des populations issues de l'approche surface

Les taux bruts d'incidence estimés à l'aide des populations calculées à partir de données télédétection apparaissent dans la figure 5, de A à E pour la méthode du pixel bâti-pondérée, et de F à J pour la méthode de la densité du bâti pondérée. Par ailleurs, la légende de ces cartes correspond à celle du taux de référence du LNH (fig. 5-4).

De manière globale, nous observons une distribution différente dans chacune de deux méthodes. D'un côté, pour les taux bruts estimés à partir de la MPBP, les valeurs d'incidence les plus élevées se distribuent sur toute la ville, même si elles se localisent, plutôt, sur les IRIS se situant sur la deuxième couronne ainsi qu'à Planoise. De l'autre, pour les taux bruts estimés à partir de la MDBP, les valeurs d'incidence les plus élevées se localisent quasiment toutes sur les IRIS se situant sur la deuxième couronne, ainsi que sur la troisième couronne de la ville (excepté à Chailluz et à Planoise).

Quant à la distribution du taux d'incidence du LNH à l'intérieur de chaque méthode de l'approche surface (tab. 5-3), nous pouvons dégager les éléments suivants.

Pour la MPBP, le premier intervalle (< 7,5) concerne un nombre relativement homogène d'IRIS dans les cinq types de bâti. Le deuxième intervalle (entre 7,5 et 11,5) présente les effectifs les plus hétérogènes : d'une part, le plus bas avec la classification par régression logistique (fig. 5-3E) ; d'autre part, le plus élevé avec la classification hiérarchisée (fig. 5-3C). Le troisième intervalle (entre 11,5 et 14,5) ne concerne qu'un faible nombre d'IRIS. Le dernier intervalle (> 14,5) est plutôt homogène et regroupe les effectifs d'IRIS les plus nombreux.

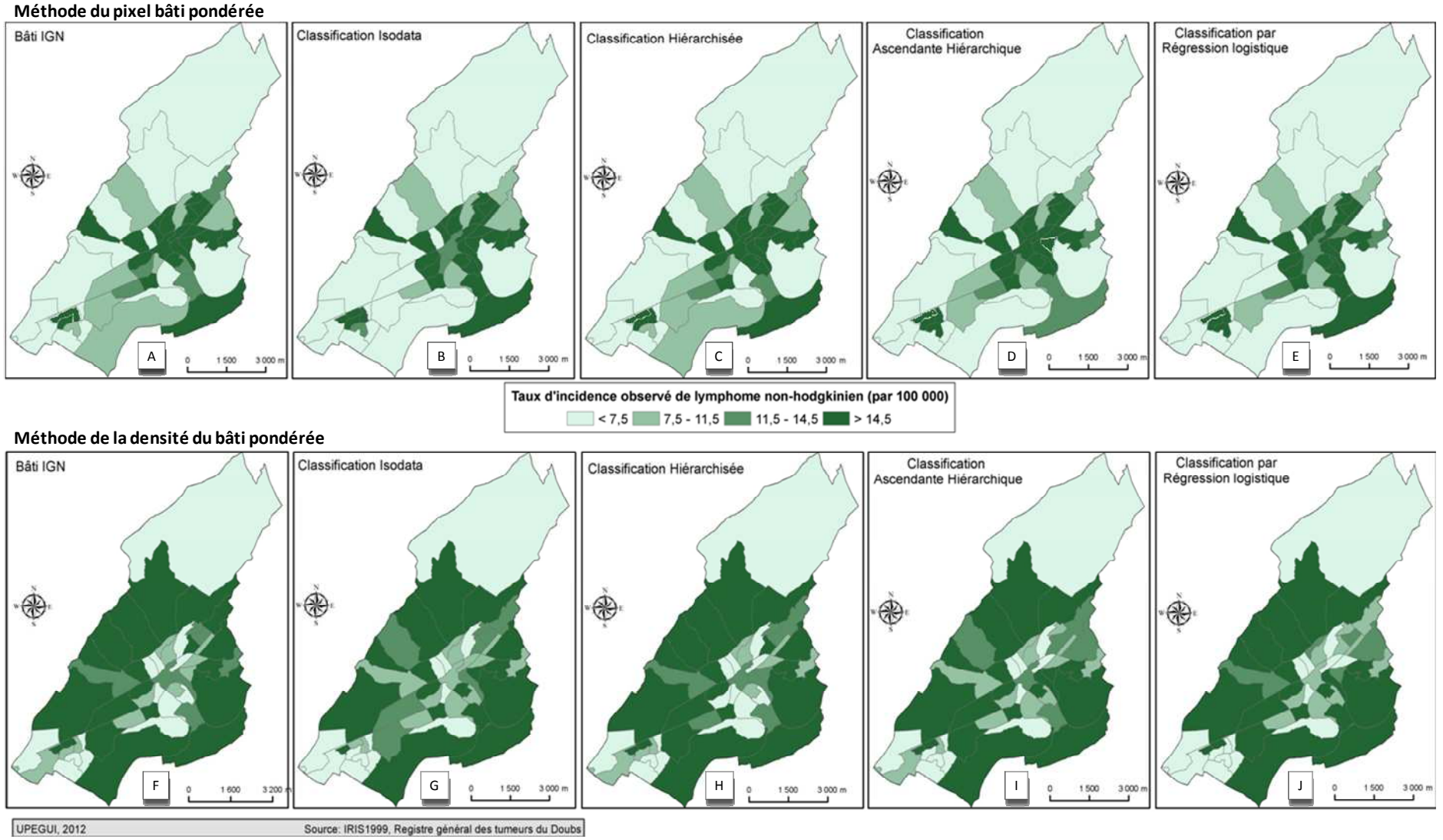
Pour la MDBP, la distribution des fréquences suit quasiment la même tendance que celle de la méthode MPBP. Toutefois, dans le tableau 5-de distribution (tab 3), elle présente les effectifs les plus élevés dans les deuxième (entre 7,5 et 11,5) et troisième (entre 11,5 et 14,5) intervalles ; et des effectifs moindres dans le dernier intervalle (> 14,5).

Tableau 5-3 : Fréquence des IRIS par méthode d'estimation de la population et par intervalle du taux du lymphome non hodgkinien

		Population issue de la méthode du pixel-bâti pondérée			
Intervalle du taux d'incidence du lymphome non hodgkinien		**< 7,5**	**7,5 – 11,5**	**11,5 – 14,5**	**> 14,5**
Source du bâtiment	Bâti IGN	15	8	5	24
	Classification ISODATA	17	6	3	26
	Classification Hiérarchisée	15	9	3	25
	Classification Ascendante Hiérarchique	16	6	5	25
	Classification Régression logistique	17	5	5	25

		Population issue de la méthode de la densité du bâti pondérée			
Intervalle du taux d'incidence du lymphome non hodgkinien		**< 7,5**	**7,5 – 11,5**	**11,5 – 14,5**	**> 14,5**
Source du bâtiment	Bâti IGN	14	8	7	20
	Classification ISODATA	13	14	8	17
	Classification Hiérarchisée	15	9	9	19
	Classification Ascendante Hiérarchique	13	13	7	19
	Classification Régression logistique	16	11	5	20

Figure 5-5 : Taux d'incidence estimés du lymphome non-hodgkinien, fondés sur la population modélisée par l'approche surface. De A à E par la méthode MPBP ; de F à J par la méthode MDBP



5.2.2.2 Estimation des taux à partir des populations issues de l'approche volume

La figure 5-6 englobe les cartes des taux bruts d'incidence calculés en utilisant les différentes populations estimées par l'approche volume (de A à E par la méthode MPBP-3D, et de F à J par la méthode MDBP-3D). Par analogie avec les estimations réalisées avec l'approche surface, la légende de ces cartes reste inchangée.

Sur ces cartes, on observe encore, globalement, une tendance différente entre les deux méthodes. Concernant la distribution spatiale des taux bruts estimés à partir de la MPBP-3D, celle-ci reste aléatoire, surtout pour les taux calculés à partir du bâti issu de la classification ascendante hiérarchique (fig. 5-6D), ainsi que celui provenant de la classification par régression logistique (fig. 5-6E). Toutefois, les estimations effectuées sur le bâti IGN (fig. 5-6A), le bâti issu de la classification ISODATA (fig. 5-6B), ainsi que celui extrait par classification hiérarchisée (fig. 5-6C) montrent une concentration des taux d'incidence élevés, d'une part, à l'ouest de la ville sur la deuxième couronne, en particulier sur la frontière entre la colline et la plaine ; et d'autre part, à Planoise. Quant aux taux bruts estimés avec la MDBP-3D, leur distribution spatiale situe les taux d'incidence hauts sur une large étendue de la ville, excluant principalement : le centre-ville, les Chaprais, Chailluz et Planoise.

Après analyse plus fine de la distribution des taux bruts d'incidence du LNH, à l'intérieur de chaque méthode (tab. 5-4), on peut faire ressortir différents éléments :

Du côté de la MPBP, le premier intervalle (< 7,5) présente un nombre d'IRIS homogène dans les cinq types de bâti. Le deuxième intervalle (entre 7,5 et 11,5) et le troisième intervalle (entre 11,5 et 14,5) ont des effectifs similaires, à l'exception de la classification par régression logistique. Le dernier intervalle (> 14,5) est relativement homogène avec les nombres d'IRIS les plus importants par rapport aux autres intervalles.

Du côté de la MDBP, le premier intervalle (< 7,5) et le deuxième intervalle (entre 7,5 et 11,5) ont des effectifs relativement homogènes entre les cinq types de bâti, bien qu'inférieurs dans le deuxième intervalle. Le troisième intervalle (entre 11,5 et 14,5) regroupe les effectifs les plus faibles, oscillant entre 2 et 6. En revanche, le dernier intervalle (> 14,5) les nombres d'IRIS les plus élevés, variant entre 19 et 24.

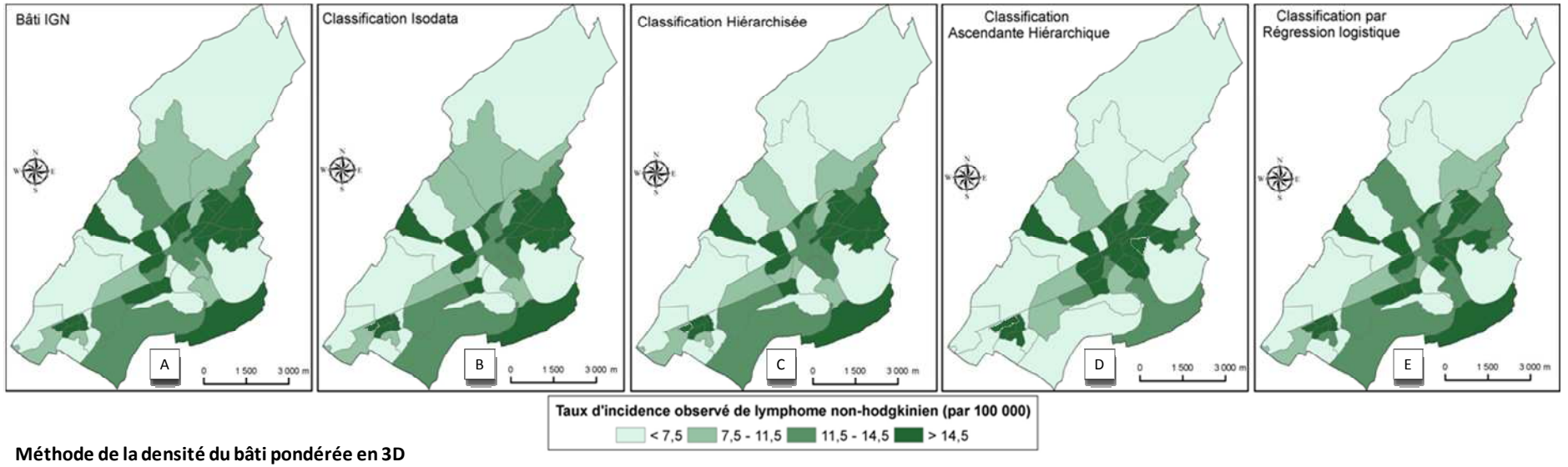
Tableau 5-4 : Fréquence des IRIS par méthode d'estimation de la population et par intervalle du taux du lymphome non hodgkinien

		Population issue de la méthode du pixel-bâti pondérée en 3D			
Intervalle du taux d'incidence du lymphome non hodgkinien		**< 7,5**	**7,5 – 11,5**	**11,5 – 14,5**	**> 14,5**
Source du bâtiment	Bâti IGN	13	8	8	23
	Classification ISODATA	13	9	9	21
	Classification Hiérarchisée	13	9	10	20
	Classification Ascendante Hiérarchique	13	10	10	19
	Classification Régression logistique	13	6	13	20

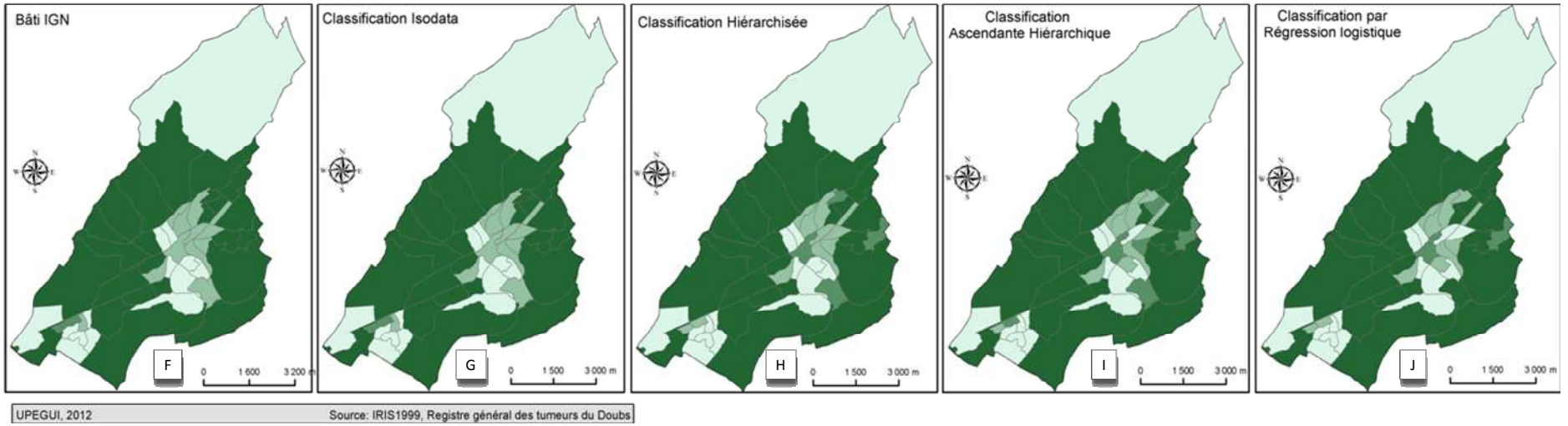
		Population issue de la méthode de la densité du bâti pondérée en 3D			
Intervalle du taux d'incidence du lymphome non hodgkinien		**< 7,5**	**7,5 – 11,5**	**11,5 – 14,5**	**> 14,5**
Source du bâtiment	Bâti IGN	14	12	2	24
	Classification ISODATA	15	11	3	23
	Classification Hiérarchisée	15	11	4	22
	Classification Ascendante Hiérarchique	15	10	6	21
	Classification Régression logistique	16	11	2	23

Figure 5-6 : Taux d'incidence estimés du lymphome non-hodgkinien, fondés sur la population modélisée par l'approche volume. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D

Méthode du pixel bâti pondérée en 3D



Méthode de la densité du bâti pondérée en 3D



5.3 Validation des estimations

5.3.1 Analyse des résultats

La validation des résultats obtenus lors du calcul des taux d'incidence (utilisant les populations estimées -cf. 4.2- comme dénominateur) s'est effectuée par le CCI, à l'aide des taux de références observés (utilisant le recensement comme dénominateur) tant pour le cancer du sein (tab. 5-5) que pour le lymphome non-hodgkinien (tab. 5-6). Toutefois, une des limites de notre travail réside dans la différence entre les dates d'acquisition des données : médicales, censitaires, satellitaires, et références, notamment la BD Topo® de l'IGN. Cette chronologique entraîne une erreur qu'on ne peut pas maitriser, même si elle est probablement modérée pour les raisons suivantes. La population de Besançon est restée stable au cours de deux dernières décennies. De plus, les nouvelles constructions édifiées, entre la date de la dernière mise à jour de la BD Topo® et celle des prises de vues des images THRS, ont été identifiées, et donc elles ont été prises en compte dans la modélisation de la population. Par ailleurs, des diagrammes bi-variés -taux estimés *vs* taux observés- (fig.5-7 à fig. 5-10) ont également été réalisés pour les deux types de cancer et pour les deux types d'approche de la population.

En ce qui concerne le cancer du sein (tab. 5-5), après analyse des CCIs, différentes informations peuvent être tirées.

D'une part, les taux estimés à partir des populations issues de la méthode du pixel bâti pondérée (MPBP) présentent des valeurs du CCI supérieures (entre 0,26 et 0,60) aux estimations de population issues de la méthode de la densité du bâti pondérée (entre 0,18 et 0,45).

D'autre part, l'utilisation des données de hauteur (c'est-à-dire l'approche « volume ») a amélioré les valeurs du CCI pour MPBP (passant d'un intervalle de valeurs [0,36 - 0,35] dans l'approche « surface », à un intervalle de valeurs [0,49 - 0,60] dans l'approche « volume »). Mais ces données n'ont pas produit le même effet sur les valeurs du CCI pour MDBP : ces valeurs du CCI ont diminué, passant d'un intervalle de valeurs entre 0,31 et 0,45 pour l'approche « surface » à un intervalle de valeurs entre 0,18 et 0,22 pour l'approche « volume ».

De plus, certains taux estimés à partir des populations calculées par MDBP-3D ne présentent d'ICC statistiquement significatifs. Il s'agit de ceux qui ont intégré les bâtis issus des classifications : ISODATA, hiérarchisée, ainsi que par régression logistique. En revanche, tous les autres taux fondés sur les populations estimées par les autres méthodes (MPBP, MPBP-3D et MDBP) utilisant les différents bâtis (issus des quatre classifications ainsi que celui de la BD Topo®), conduisent à un ICC statistiquement significatifs ($p < 0,05$).

Par ailleurs, dans tous les cas, les taux estimés à partir des bâtis issus des traitements télédétections (classification fondée sur le pixel ou sur l'objet) donnent des valeurs du CCI supérieures (ou égales par MDBP-3D) à celles obtenues en utilisant le bâti de référence, c'est-à-dire celui de l'IGN.

Tableau 5-5 : Coefficient de corrélation intra-classe entre les taux des cancers du sein estimés et les taux des cancers du sein observés, pour les IRIS : N=52

		Taux bruts d'incidence observés vs taux bruts d'incidence estimés à partir de			
		Méthode du pixel-bâti pondérée CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,26 (-0,01-0,49)	0,03	0,31 (0,04-0,54)	0,01
	Classification ISODATA	0,35 (0,08-0,56)	0,005	0,38 (0,12-0,59)	0,002
	Classification Hiérarchisée	0,31 (0,04-0,53)	0,01	0,32 (0,05-0,55)	0,008
	Classification Ascendante Hiérarchique	0,30 (0,03-0,53)	0,01	0,45 (0,20-0,64)	0,0003
	Classification Régression logistique	0,31 (0,04-0,53)	0,01	0,37 (0,11-0,58)	0,003
		Méthode du pixel-bâti pondérée en 3D CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée en 3D** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,49 (0,25-0,67)	0,0001	0,18 (-0,02-0,49)	0,10
	Classification ISODATA	0,59 (0,38-0,74)	1,8e-06	0,18 (-0,09-0,43)	0,09
	Classification Hiérarchisée	0,57 (0,36-0,73)	3,9e-06	0,19 (-0,08-0,44)	0,08
	Classification Ascendante Hiérarchique	0,52 (0,28-0,69)	3,8e-05	0,22 (-0,05-0,47)	0,05
	Classification Régression logistique	0,60 (0,39-0,75)	9,4e-07	0,19 (-0,08-0,44)	0,08

S'agissant du lymphome non-hodgkinien (tab. 5-6), l'analyse des résultats du calcul du CCI permet de déduire plusieurs éléments.

Tableau 5-6 : Coefficient de corrélation intra-classe entre les taux des lymphomes non-hodgkiniens estimés et les taux des lymphomes non-hodgkiniens observés pour les IRIS : N=52

		Taux bruts d'incidence observés vs taux bruts d'incidence estimés à partir de			
		Méthode du pixel-bâti pondérée CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,29 (0,02-0,52)	0,02	0,49 (0,25-0,67)	0,0001
	Classification ISODATA	0,33 (0,06-0,54)	0,008	0,63 (0,43-0,77)	2,3e-07
	Classification Hiérarchisée	0,30 (0,03-0,53)	0,01	0,52 (0,29-0,69)	3,6e-05
	Classification Ascendante Hiérarchique	0,28 (0,00-0,51)	0,02	0,67 (0,49-0,80)	1,4e-08
	Classification Régression logistique	0,28 (0,01-0,51)	0,02	0,66 (0,48-0,79)	3,4e-08
		Méthode du pixel-bâti pondérée en 3D CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée en 3D** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,55 (0,33-0,71)	9,2e-06	0,28 (0,01-0,51)	0,02
	Classification ISODATA	0,63 (0,43-0,77)	2,3e-07	0,29 (0,02-0,52)	0,02
	Classification Hiérarchisée	0,54 (0,32-0,71)	1,4e-05	0,32 (0,06-0,55)	0,008
	Classification Ascendante Hiérarchique	0,52 (0,29-0,69)	3,0e-05	0,37 (0,10-0,58)	0,003
	Classification Régression logistique	0,56 (0,35-0,72)	5,1e-06	0,33 (0,07-0,55)	0,007

En premier lieu, les taux estimés utilisant soit les populations issues de MPBP soit celles issues de MDBP présentent globalement des valeurs du CCI semblables, à savoir : entre 0,28 et 0,63 pour MPBP ; entre 0,28 et 0,67 pour MDBP.

En deuxième lieu, l'utilisation des données de hauteur (c'est-à-dire l'approche « volume ») produit le même effet sur les estimations que dans le cancer du sein. Ainsi, les valeurs du CCI pour MPBP augmentent, passant d'un intervalle de valeurs [0,28 - 0,33] dans

l'approche « surface » à un intervalle de valeurs [0,52 - 0,63] dans l'approche « volume » ; tandis que celles pour MDBP diminuent (passant d'un intervalle de valeurs [0,49 - 0,67] pour l'approche « surface » à un intervalle de valeurs [0,28 - 0,33] pour l'approche « volume »).

En troisième lieu, toutes les estimations sont statistiquement significatives ($p < 0,05$).

Par ailleurs, l'analyse des graphiques bi-variés permet de porter les conclusions suivantes.

Premièrement, les taux estimés par MPBP sont globalement surestimés, que ce soit pour le cancer du sein (fig. 5-7 de A à E) ou pour le lymphome non-hodgkinien (fig. 5-9 idem). Cela peut être déduit du grand nombre de points se localisant au-dessus de la diagonale. Toutefois, cet effet est atténué dans le cas du lymphome non-hodgkinien : il existe également une sous-estimation des taux hauts d'incidence, matérialisée par les points du dernier quartile (losanges verts) localisés en dessous et sur la droite de la diagonale.

Deuxièmement, l'injection de la hauteur dans la MPBP (c'est-à-dire le MPBP-3D) permet un meilleur ajustement des points sur la diagonale. Il s'ensuit qu'on observe un écart plus faible entre les points et la diagonale (par rapport à la représentation des taux estimés par MPBP), tant pour le cancer du sein (fig. 5-8 de A à E) que pour le lymphome non-hodgkinien (fig. 5-10 de A à E).

Troisièmement, de même que pour les taux estimés par MPBP, les graphiques correspondant aux taux estimés par MDBP (fig. 5-7 de F à J pour le cancer du sein, et fig. 5-9 de F à J pour le lymphome non-hodgkinien) mettent en évidence la surestimation des taux bruts d'incidence. Cependant, pour le lymphome non-hodgkinien, l'ajustement des points sur la diagonale est meilleur que pour le cancer du sein.

Quatrièmement, l'injection de la hauteur est défavorable pour la méthode de la densité du bâti pondérée, que ce soit pour le cancer du sein (fig. 5-8 de F à J) ou pour le lymphome non-hodgkinien (fig. 5-10 idem). Dans les deux cas, on observe une surestimation des taux bruts d'incidence (bas et haut). Cela s'observe également sur les cartes des taux d'incidence (fig. 5-3 de F à J pour le cancer du sein, et fig. 5-6 de F à J pour le lymphome non-hodgkinien) où le dernier quartile occupe une grande partie de l'étendue de la ville.

Figure 5-7 : Diagrammes bi-variés des taux estimés à partir des populations produites par l'approche surface *vs* les taux observés des cancers du sein. De A à E par la méthode MPBP ; de F à J par la méthode MDBP

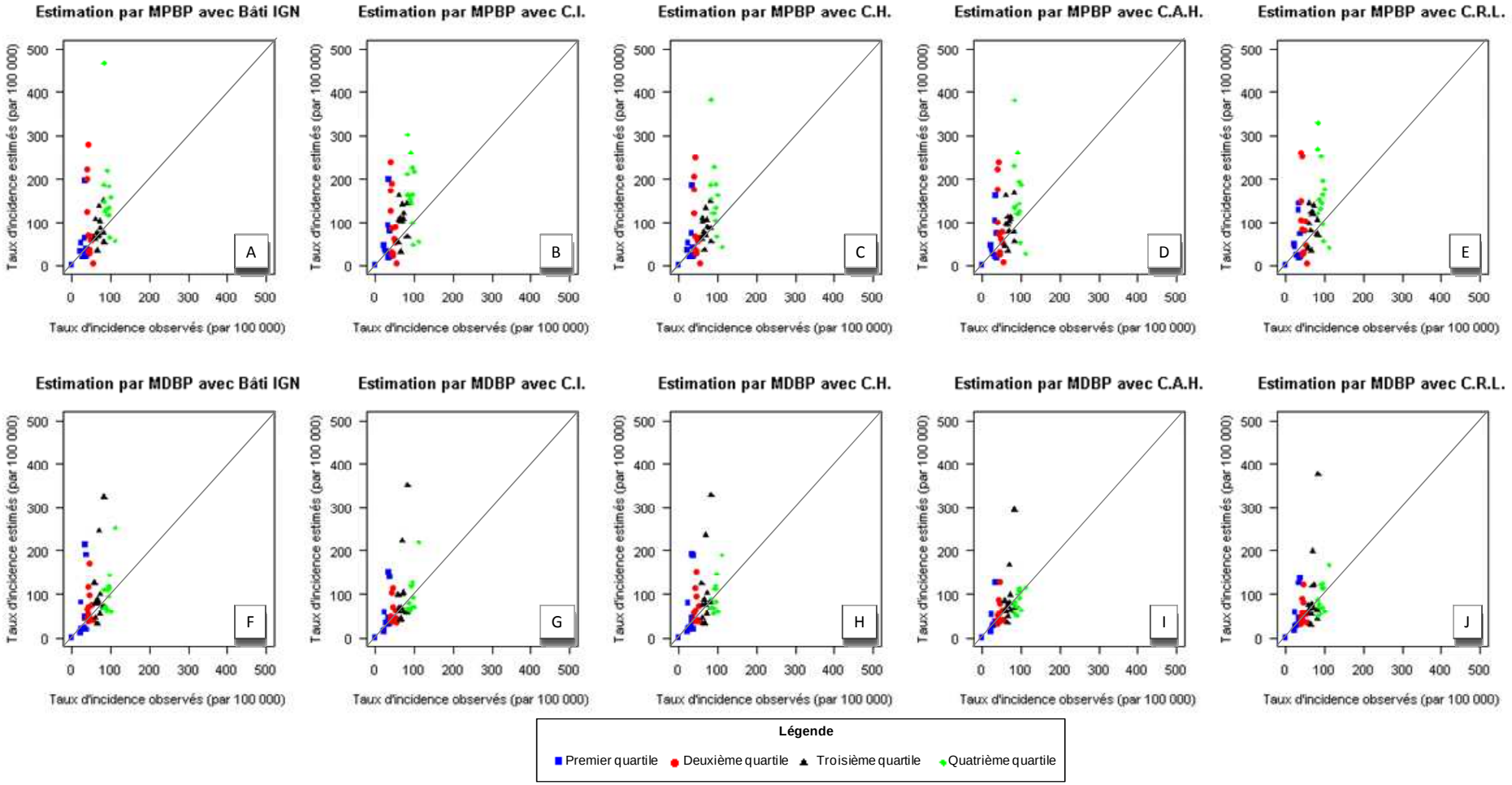


Figure 5-8 : Diagrammes bi-variés des taux estimés à partir des populations produites par l'approche volume *vs* les taux observés des cancers du sein. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D

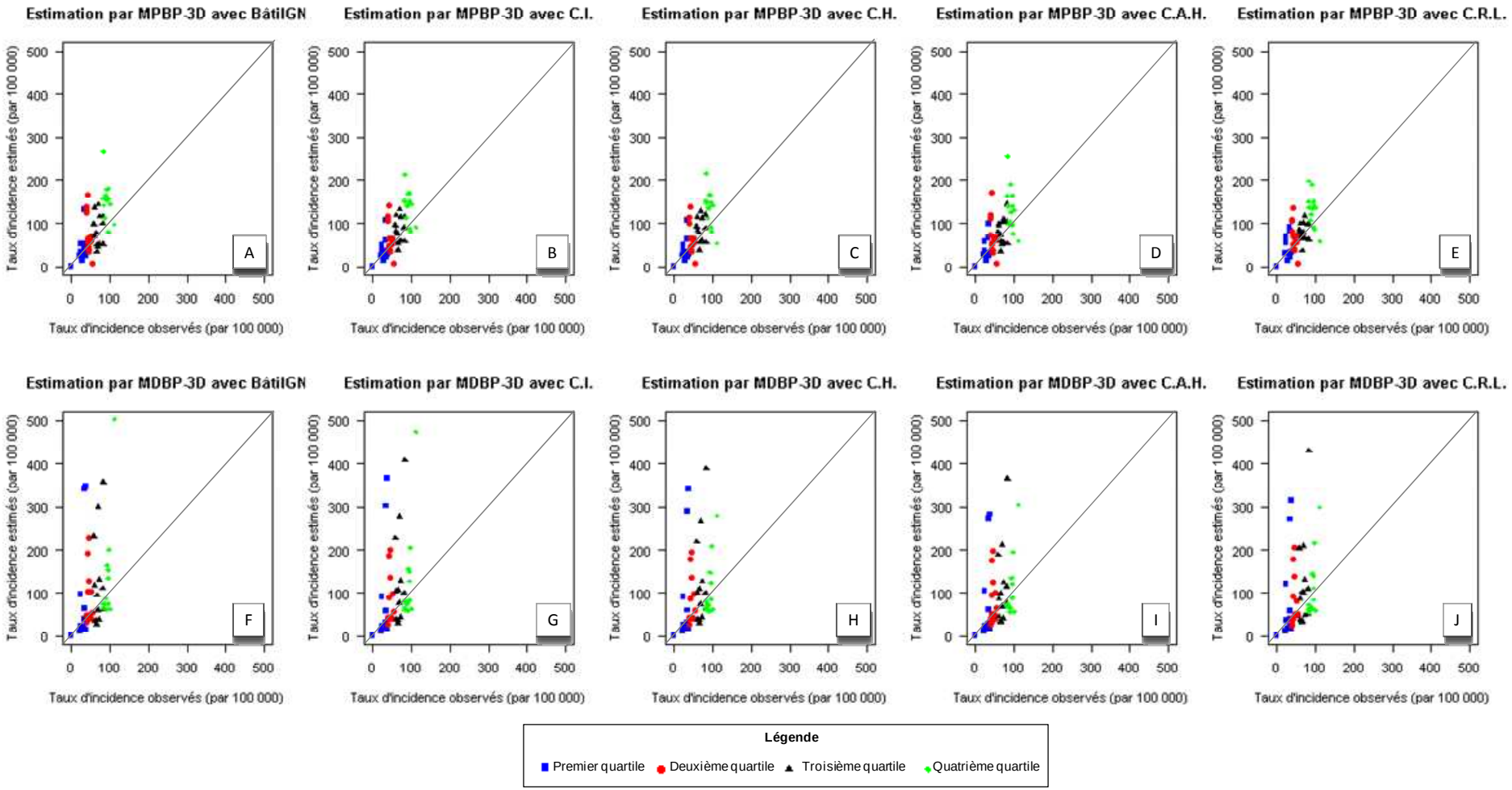


Figure 5-9 : Diagrammes bi-variés des taux estimés à partir des populations produites par l'approche surface *vs* les taux observés des lymphomes non-hodgkiniens. De A à E par la méthode MPBP ; de F à J par la méthode MDBP

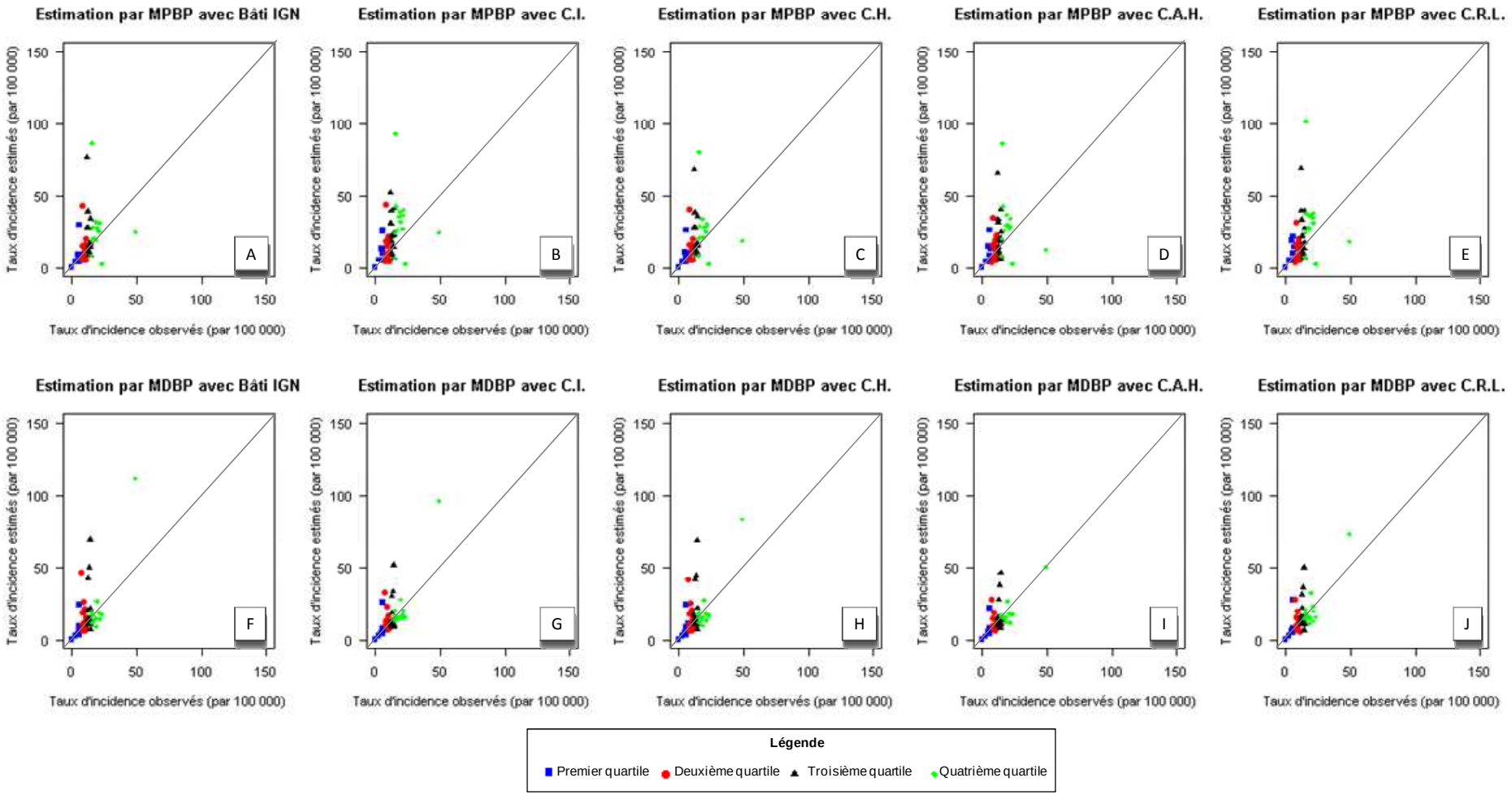
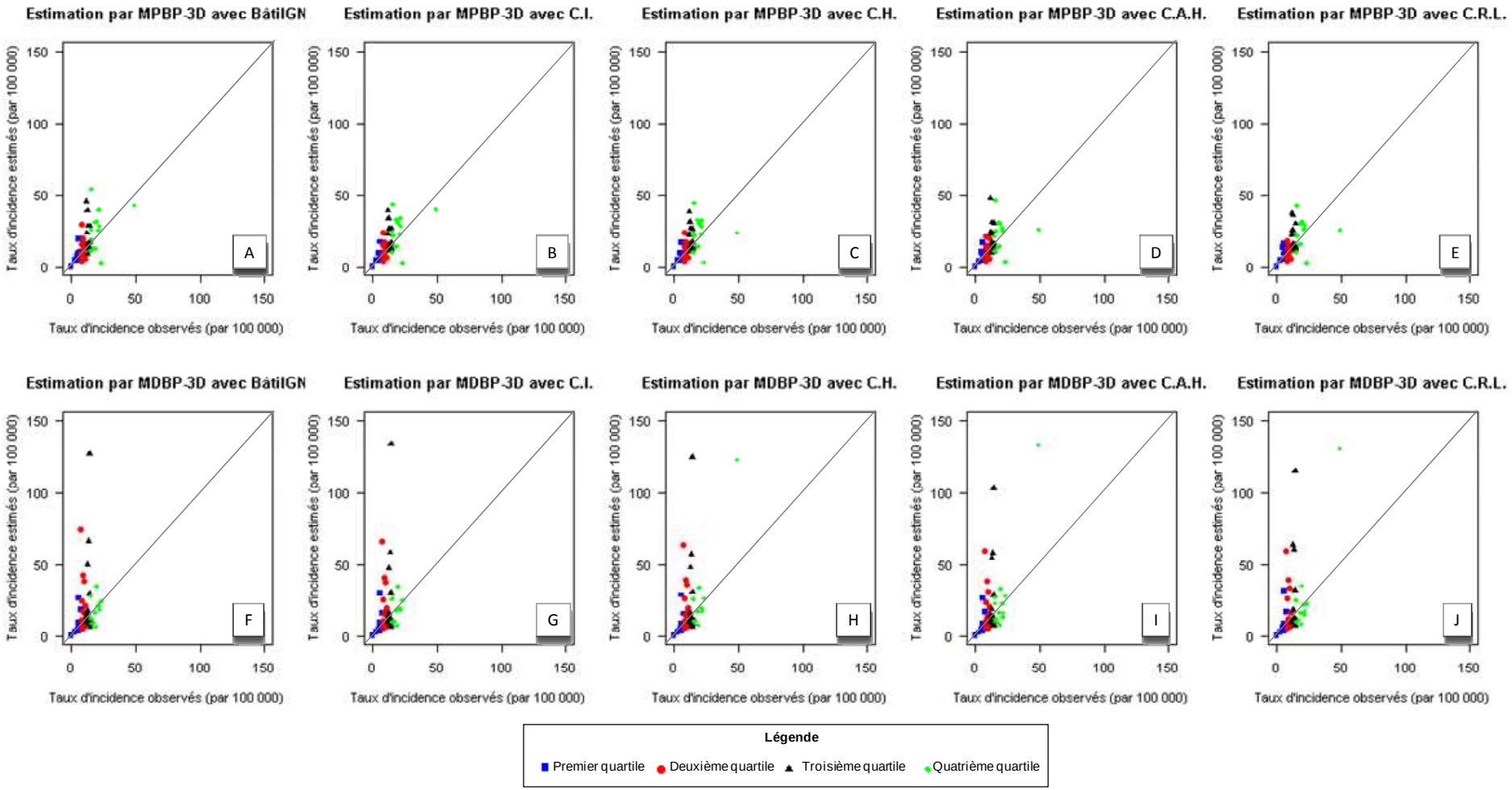


Figure 5-10 : Diagrammes bi-variés des taux estimés à partir des populations produites par l'approche volume *vs* les taux observés des lymphomes non-hodgkiniens. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D



Compte tenu des résultats obtenus, deux conclusions préliminaires peuvent être tirées.

D'une part, les résultats ne sont pas identiques entre le cancer du sein -maladie plus courante- et le lymphome non-hodgkinien -maladie plutôt rare.

D'autre part, globalement, les meilleurs résultats s'obtiennent avec l'utilisation des populations estimées par MPBP-3D, c'est-à-dire par l'indicateur le plus simple -somme des pixels bâtis-couplé à la hauteur (approche « volume »).

5.3.2 Analyse de sensibilité

Suite aux résultats obtenus, trois tests ont été réalisés afin de prouver la sensibilité de la méthode, dont le premier nous permet de comparer nos résultats à ceux obtenus par Viel et Tran (2009).

5.3.2.1 L'influence des incidences nulles

Notre objectif est d'identifier l'influence des incidences nulles observées sur l'estimation des taux bruts d'incidence quand on utilise les populations dérivées de la modélisation à partir des données télédétection. Pour ce faire, le CCI a été calculé entre les taux observés et les taux calculés, cette fois-ci en excluant les IRIS où aucun nouveau cas n'avait été enregistré (par le Registre général des tumeurs du Doubs) pendant la période observée. Quant aux populations, nous avons repris celles obtenues initialement par les 52 IRIS, c'est-à-dire sans aucune modification résultant des analyses de sensibilité effectuées.

Pour le cancer du sein, 3 IRIS ont été exclus : Lafayette, Châteaufarine, et Chailluz. Le tableau 5-7 présente les résultats du calcul du CCI entre les taux estimés et taux observés. De l'analyse de ces valeurs nous pouvons remarquer, tout d'abord et globalement, une diminution de toutes les valeurs du CCI. Par exemple, les valeurs du CCI basés sur la modélisation MBPP-3D sont passées d'un intervalle [0,49 - 0,60] en utilisant les 52 IRIS, à un intervalle [0,40 - 0,51], en utilisant seulement les IRIS avec au moins un cas. En revanche, l'ordre de performance des méthodes employées pour estimer la population se maintient. Ainsi, le classement des résultats obtenus (selon une signification statistique décroissante) est le suivant : MPBP-3D, MDBP, MPBP, puis MDBP-3D (avec des CCIs non statistiquement significatifs - $p > 0,05$).

Tableau 5-7 : Coefficient de corrélation intra-classe entre les taux des cancers du sein estimés et le taux des cancers du sein observés pour les 49 IRIS ayant présenté au moins un nouveau cas

		Taux bruts d'incidence observés vs taux bruts d'incidence estimés à partir de			
		Méthode du pixel-bâti pondérée CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,19 (-0,10-0,44)	0,09	0,22 (-0,06-0,47)	0,06
	Classification ISODATA	0,27 (-0,01-0,51)	0,03	0,30 (0,02-0,53)	0,02
	Classification Hiérarchisée	0,23 (-0,05-0,48)	0,05	0,23 (-0,05-0,48)	0,05
	Classification Ascendante Hiérarchique	0,23 (-0,05-0,47)	0,06	0,35 (0,08-0,57)	0,006
	Classification Régression logistique	0,23 (-0,05-0,48)	0,05	0,28 (0,00-0,52)	0,02
		Méthode du pixel-bâti pondérée en 3D CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée en 3D** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,40 (0,14-0,61)	0,001	0,12 (-0,16-0,39)	0,20
	Classification ISODATA	0,50 (0,27-0,69)	8,3e-05	0,13 (-0,16-0,39)	0,19
	Classification Hiérarchisée	0,49 (0,24-0,67)	1,7e-04	0,13 (-0,16-0,39)	0,19
	Classification Ascendante Hiérarchique	0,43 (0,17-0,63)	9,1e-04	0,15 (-0,13-0,41)	0,15
	Classification Régression logistique	0,51 (0,27-0,69)	6,9e-05	0,13 (-0,16-0,39)	0,19

Pour le lymphome non-hodgkinien, 6 IRIS ont été exclus de cette analyse, à savoir : Victor Hugo, la Grette, Epoisses-Champagne, ainsi que les trois déjà exclus dans le cas des cancers du sein (Lafayette, Châteaufarine, et Chailluz). Les valeurs du CCI, qui correspondent aux taux bruts estimés excluant ces IRIS apparaissent dans le tableau 5-8.

Tableau 5-8 : Coefficient de corrélation intra-classe entre les taux des lymphomes non-hodgkiniens estimés et le taux des lymphomes non-hodgkiniens observés pour les 46 IRIS, ayant présenté au moins un nouveau cas

		Taux bruts d'incidence observés vs taux bruts d'incidence estimés à partir de			
		Méthode du pixel-bâti pondérée CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,17 (-0,13-0,43)	0,13	0,42 (0,15-0,63)	0,001
	Classification ISODATA	0,20 (-0,09-0,47)	0,08	0,56 (0,33-0,73)	1,9e-05
	Classification Hiérarchisée	0,16 (-0,13-0,43)	0,14	0,43 (0,17-0,64)	0,001
	Classification Ascendante Hiérarchique	0,14 (-0,15-0,41)	0,17	0,57 (0,35-0,74)	1,1e-05
	Classification Régression logistique	0,16 (-0,13-0,43)	0,14	0,58 (0,36-0,74)	7,9e-06
		Taux bruts d'incidence observés vs taux bruts d'incidence estimés à partir de			
		Méthode du pixel-bâti pondérée en 3D CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée en 3D** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,43 (0,17-0,64)	0,001	0,24 (-0,05-0,50)	0,05
	Classification ISODATA	0,51 (0,27-0,70)	0,0001	0,25 (-0,04-0,50)	0,05
	Classification Hiérarchisée	0,38 (0,11-0,61)	0,003	0,26 (-0,03-0,51)	0,04
	Classification Ascendante Hiérarchique	0,37 (0,09-0,60)	0,005	0,31 (0,02-0,54)	0,02
	Classification Régression logistique	0,41 (0,14-0,63)	0,001	0,27 (-0,02-0,52)	0,03

L'analyse de ces résultats permet de déduire les deux points suivants. D'une part, l'exclusion des unités géographiques (IRIS) qui n'enregistrent pas de nouveaux cas dans la période d'observation diminue le degré d'association entre les taux observés et les taux estimés, comme pour le cancer du sein. Ainsi, les valeurs du CCI régressent vers un intervalle [0,37 - 0,51] par MDBP-3D au lieu de [0,52 - 0,63]. D'autre part, la performance des taux estimés selon les méthodes de modélisation de la population, est différente de celle rapportée pour le cancer du sein, ce qui vérifie les résultats obtenus avant l'analyse de

sensibilité (voir 5.2.1, tab. 5-6). C'est-à-dire que les valeurs du CCI obtenues par la méthode MDBP sont plus élevées que celles obtenues en utilisant les méthodes MPBP-3D ou MPBP, avec un intervalle [0,42 - 0,58] (même si cet intervalle est plus faible que celui obtenu en utilisant les 52 IRIS [0,49 - 0,67]).

En allant plus loin dans cette analyse, nous pouvons mettre en évidence certains faits. En premier lieu, les taux bruts d'incidence estimés à partir de populations modélisées par la méthode dasymétrique –s'appuyant sur la variable « bâti / non bâti » résultant du traitement des données télédétection- est sensible à la présence/absence de cas incidents à l'intérieur des zones géographiques qui constituent la zone étudiée. En effet, la présence des zones d'analyse (géographiques), n'ayant enregistré aucun nouveau cas (pendant la période observée), augmente la précision globale des estimations dans l'étude. En revanche, dans les zones ayant enregistré des nouveaux cas dans toutes leurs zones d'analyse, les résultats attendus lors de l'estimation des taux bruts d'incidence, sont légèrement moins performants. L'explication en est simple. Avec un numérateur égal à zéro, le taux d'incidence estimé sera toujours nul (que la population au dénominateur soit sous ou surestimée), égal au taux d'incidence observé (également nul), et augmentant donc l'ICC correspondant.
En deuxième lieu, les résultats diffèrent entre une maladie rare comme le lymphome non hodgkinien (222 nouveaux cas observés en 16 ans), et une maladie plus fréquente comme le cancer du sein (491 nouveaux cas observés en 7ans). Les deux distributions suivent une loi de Poisson[56] mais la fluctuation aléatoire est beaucoup plus importante dans le cas des petits effectifs. Elle vient donc se surajouter, pour les lymphomes, à toutes les autres causes d'imprécision ou d'erreurs. Pour diminuer cette fluctuation aléatoire et obtenir une puissance statistique suffisante[57], différents auteurs ont cumulé les effectifs (Maroko *et al*.,

[56] Selon l'Encyclopedia of Biostatistics (1998) : Pour que les données suivent une distribution de Poisson trois conditions doivent être réunies : 1) Dans n'importe quel intervalle très petit (plus petit par exemple qu'une milliseconde ou un nanomètre) la probabilité d'une occurrence de l'événement est proportionnelle à la taille de l'intervalle. 2) La probabilité que l'intervalle contienne deux événements ou plus devient plus petite quand l'intervalle devient plus petit, au point que, pour les buts pratiques, elle est ignorée. 3) Ce qui se produit dans un petit intervalle quelconque est indépendant de ce qui se produit dans un autre petit intervalle, qui ne recouvre pas le premier. Ces conditions impliquent que les événements se produisent au-dessus du temps ou de l'espace à un taux constant en moyenne, chaque événement se produisant indépendamment et au hasard. Ainsi, quand *n* augmente indéfiniment, la probabilité binomiale définit la distribution de probabilité de Poisson.

[57] Probabilité d'obtenir un résultat statistiquement significatif, qui permet de conclure à l'existence de l'effet, par les méthodes employées. (Champely, 2006 ; Laurencelle, 2007)

2011 ; Viel et Tran, 2009 ; Woronoff et Danzon, 2008 ; entre autres). C'est également ce que nous avons fait dans notre étude.

Par ailleurs, si nous comparons nos résultats obtenus -lors du calcul des taux bruts d'incidence et cette analyse de sensibilité- avec ceux obtenus par Viel et Tran (2009) ayant utilisé une image LandSat, nous pouvons proposer certaines conclusions.

Les résultats obtenus –que ce soit en intégrant tous les IRIS ou en excluant ceux n'ayant pas de nouveaux cas enregistrés- avec la classification ISODATA couplée à l'approche surface (en particulier avec MDBP) pour le calcul des taux bruts d'incidence des cancers du sein, sont semblables, que ce soit avec une image THRS (GeoEye) ou avec une image HRS (LandSat). En effet, les valeurs du CCI ont été : d'une part, pour les 52 IRIS, 0,38 (THRS) *vs* 0,40 (HRS) ; d'autre part, pour les 49 IRIS, 0,30 (THRS) *vs* 0,31 (HRS).

Les valeurs du CCI obtenues en, utilisant le bâti extrait par classification hiérarchisée (CH) -qui s'est révélée la plus performante lors de nos analyses- couplée avec l'approche volume (en particulier avec MPBP) pour le calcul des taux bruts d'incidence des cancers du sein sont plus élevés que celles obtenues avec ISODATA. Aussi, ces valeurs passent-elles : d'un côté, pour les 52 IRIS, de 0,38 (ISODATA) à 0,57 (CH) ; de l'autre, pour les 49 IRIS, de 0,30 (ISODATA) à 0,49 (CH).

En outre, les conclusions précédentes ne sont pas applicables au lymphome-non hodgkinien, car les valeurs obtenues avec l'image THRS sont moindres que celles obtenues avec l'image HRS. Ainsi, nous avons obtenu pour les 52 IRIS (avec ISODATA et MDBP) une valeur du CCI égale à 0,63 tandis que Viel et Tran (2009) ont obtenu une valeur du CCI égale à 0,73. D'ailleurs, la même valeur du CCI égale à 0,63 a été obtenue en utilisant CH avec MPBP-3D. Quant aux valeurs obtenues en excluant les IRIS ayant les cas d'incidents « nuls », elles sont de : 0,65 pour ISODATA et MDBP sur HRS ; 0,56 pour ISODATA et MPBP-3D sur THRS ; et 0,38 pour CH et MPBP-3D sur THRS.

5.3.2.2 L'influence de la surestimation de la population due a une utilisation non résidentielle des sols

Pour établir l'impact de la surestimation des populations -provoquée par les zones ayant une utilisation des sols essentiellement non résidentielle- sur l'estimation des taux bruts d'incidence, nous avons exclu les IRIS à activité commerciale et industrielle (Châteaufarine et les Tilleroyes - cf. 4.3.2, tableau 4-3).

Tableau 5-9 : Coefficient de corrélation intra-classe entre les taux des cancers du sein estimés et les taux des cancers du sein observés pour les IRIS, excluant Châteaufarine et les Tilleroyes : N=50

		Taux bruts d'incidence observés vs taux bruts d'incidence estimés à partir de			
		Méthode du pixel-bâti pondérée CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,27 (-0,00-0,50)	0,03	0,29 (0,01-0,52)	0,02
	Classification ISODATA	0,37 (0,11-0,59)	0,003	0,36 (0,09-0,58)	0,004
	Classification Hiérarchisée	0,32 (0,05-0,55)	0,01	0,30 (0,03-0,53)	0,01
	Classification Ascendante Hiérarchique	0,31 (0,04-0,54)	0,01	0,43 (0,17-0,63)	8,9e-04
	Classification Régression logistique	0,32 (0,06-0,55)	0,009	0,35 (0,08-0,57)	0,005
		Méthode du pixel-bâti pondérée en 3D CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée en 3D** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,51 (0,27-0,69)	6,1e-05	0,16 (-0,12-0,41)	0,13
	Classification ISODATA	0,62 (0,42-0,76)	5,9e-07	0,17 (-0,11-0,42)	0,12
	Classification Hiérarchisée	0,60 (0,38-0,75)	1,8e-06	0,17 (-0,11-0,43)	0,11
	Classification Ascendante Hiérarchique	0,53 (0,30-0,70)	2,8e-05	0,20 (-0,08-0,45)	0,07
	Classification Régression logistique	0,63 (0,43-0,77)	3,5e-07	0,17 (-0,11-0,43)	0,11

En ce qui concerne le cancer du sein (tab. 5-9), si on compare ces valeurs du CCI obtenues en excluant les zones non résidentielles avec les résultats obtenus sans exclure aucun IRIS (tab. 5-5), nous observons que les changements sont très faibles, à savoir, entre 0,01 et 0,03. Prenons comme exemple le bâti de l'IGN : la valeur du CCI, entre les taux bruts d'incidence observés et ceux calculés à partir de la population estimée par MPBP, passe de 0,26 (pour tous les IRIS) à 0,27 (excluant les IRIS commerciaux et industriels) ; pour MDBP, cette variation est de 0,31 (avec tous les IRIS) à 0,29 (en excluant deux IRIS) ; pour MPBP-3D, le CCI passe de 0,49 (avec tous les IRIS) à 0,51 (en excluant deux IRIS) ; ou encore, il varie de 0,16 (tous les IRIS) à 0,18 (excluant deux IRIS) pour l'estimation effectuée par MDBP-3D.

Au-delà de ces faibles variations sur le CCI, le degré d'association entre les taux bruts d'incidence observés et ceux obtenus à travers les populations modélisées par différentes méthodes se maintient.

Tableau 5-10 : Coefficient de corrélation intra-classe entre les taux du lymphome non-hodgkinien estimés et les taux du lymphome non-hodgkinien observés pour les IRIS : N=50, sans IRIS commerciaux / industriels

		Taux bruts d'incidence observés vs taux bruts d'incidence estimés à partir de			
		Méthode du pixel-bâti pondérée CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,33 (0,06-0,55)	0,009	0,49 (0,24-0,67)	0,0001
	Classification ISODATA	0,37 (0,11-0,59)	0,003	0,63 (0,43-0,77)	2,9e-07
	Classification Hiérarchisée	0,34 (0,07-0,56)	0,007	0,51 (0,27-0,69)	5,4e-05
	Classification Ascendante Hiérarchique	0,31 (0,04-0,54)	0,01	0,67 (0,48-0,80)	4,5e-08
	Classification Régression logistique	0,32 (0,05-0,55)	0,01	0,67 (0,47-0,79)	4,9e-08
		Méthode du pixel-bâti pondérée en 3D CCI – 95%IC	**p**	**Méthode de la densité du bâti pondérée en 3D** CCI – 95%IC	**p**
Source du bâtiment	Bâti IGN	0,61 (0,40-0,76)	9,5e-07	0,28 (0,00-0,51)	0,02
	Classification ISODATA	0,69 (0,51-0,81)	1,0e-08	0,28 (0,00-0,52)	0,02
	Classification Hiérarchisée	0,59 (0,37-0,74)	2,8e-06	0,32 (0,04-0,54)	0,01
	Classification Ascendante Hiérarchique	0,57 (0,34-0,73)	7,4e-06	0,35 (0,09-0,57)	0,005
	Classification Régression logistique	0,62 (0,41-0,76)	6,7e-07	0,32 (0,06-0,55)	0,009

S'agissant du lymphome non hodgkinien (tab. 5-10), la même comparaison a été faite entre valeurs du CCI obtenues : d'une part, avec les 52 IRIS (tab. 5-6) ; et d'autre part, avec les 50 IRIS (en excluant ceux non résidentiels : tab. 5-10). Les écarts sont faibles. Pour le bâti de l'IGN, la valeur du CCI, entre les taux bruts d'incidence observés et ceux calculés à partir de la population estimée par MPBP, passe de 0,29 (52 IRIS) à 0,33 (50 IRIS) ; pour MPBP-3D,

la variation est de 0,55 (52 IRIS) à 0,61 (50 IRIS) ; tandis que pour MDBP et pour MDBP-3D, les valeurs ne varient pas, restant à 0,49 et à 0,28 respectivement.

Ainsi, la seule conclusion additionnelle que nous pouvons déduire de cette analyse de sensibilité est que la surestimation de la population en raison d'une utilisation non résidentielle n'a presque aucune influence sur le calcul des taux d'incidence, probablement parce que l'analyse de sensibilité n'a consisté qu'en l'exclusion de deux IRIS.

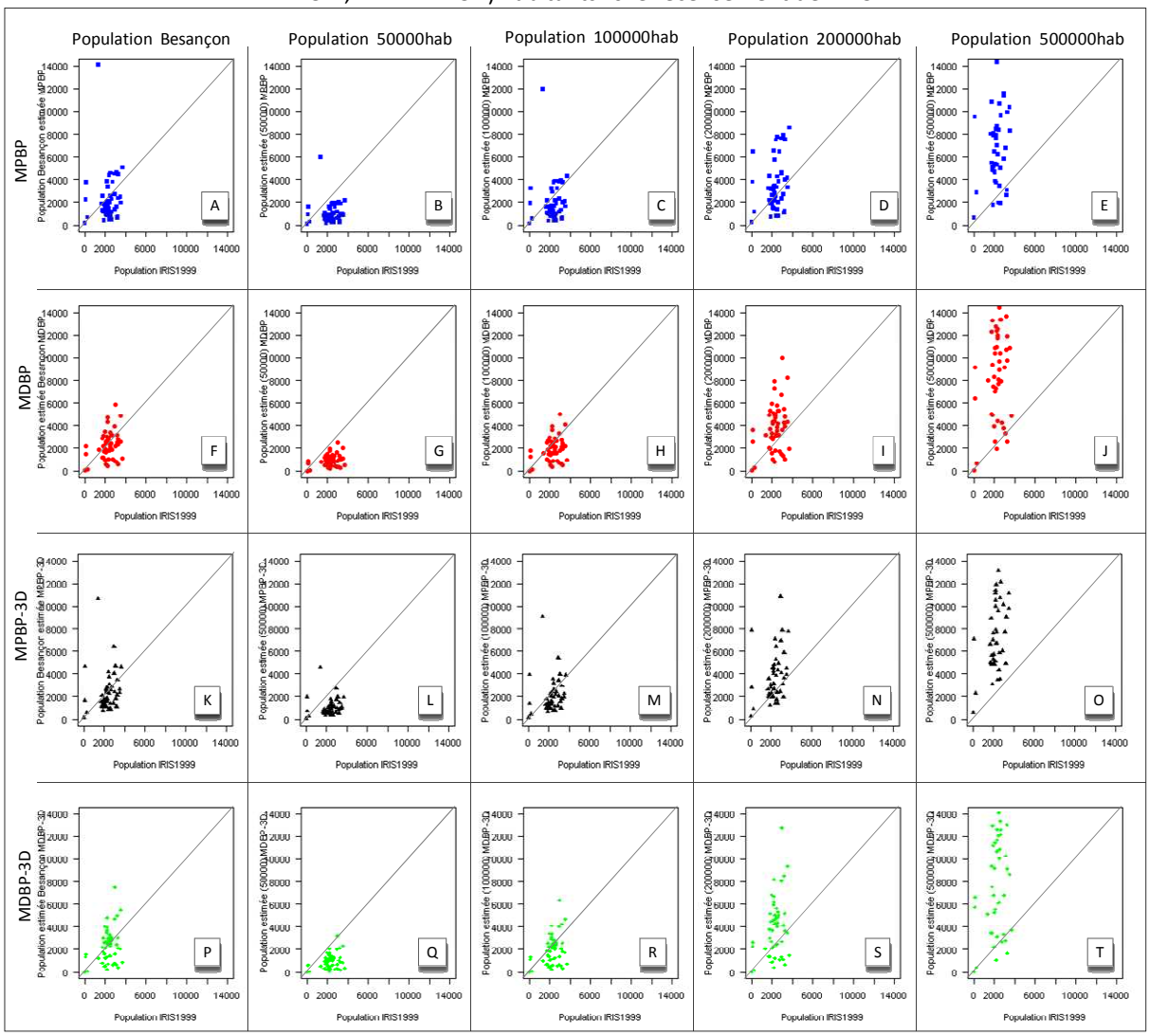
5.3.2.3 L'impact du paramètre global

Une dernière analyse de sensibilité a été effectuée sur notre méthode, visant à établir l'importance du paramètre « global », c'est-à-dire la population totale de la zone étudiée, dans l'identification des zones ayant un taux brut d'incidence élevé, pour une maladie donnée. Or, on sait en théorie que la méthode dasymétrique, (généralement plus fiable que les autres méthodes d'estimation de la population, Wu *et al.*, 2008), utilise à la fois le recensement -comme paramètre global- et, dans notre cas, la variable « bâti / non bâti » -comme source des indicateurs de répartition- pour modéliser la population (voir chapitre 4).

Sur le plan pratique, nous avons simulé quatre populations totales différentes avec : 50 000, 100 000, 200 000, et 500 000 habitants, lesquelles ont été distribuées proportionnellement à la population réelle dans chacun des IRIS, afin d'avoir une « population de référence simulée » à son niveau. Les « populations simulées » -50 000, 100 000, 200 000, et 500 000 habitants-, correspondant à la population totale de la commune (échelle de départ), ont été utilisées comme paramètre global dans l'application de la méthode dasymétrique, pour les quatre méthodes : MPBP, MDBP, MPBP-3D et MDPB-3D. C'est avec les bâtis issus de la classification hiérarchisée que cette expérience a été réalisée. Ainsi, nous avons modélisé les populations au niveau des IRIS (échelle d'arrivée). Ces populations estimées ont été corrélées : d'une part, à la population de référence de l'INSEE (fig. 5-11) ; d'autre part, aux populations de référence simulées (fig. 5-12). Dans tous les diagrammes bi-variés les limites ont été les mêmes (de 0 à 14 000) pour les différentes méthodes : MPBP (fig. 5-11 de A à E, et fig. 5-12 idem) ; MDBP (fig. 5-11 de F à J, et fig. 5-12 idem) ; MPBP-3D (fig. 5-11 de K à O, et fig. 5-12 idem) ; et MDBP (fig. 5-11 de P à T, et fig. 5-12 idem). De manière analogue, le même ordre de gauche à droite s'est maintenu dans la distribution intérieure des graphiques, comme suit : population réelle de Besançon (fig. 5-11 A, 11 F, 11 K, 11 P ; et fig. 5-12 idem) ; population simulée avec 50 000 habitants (fig. 5-11 B,

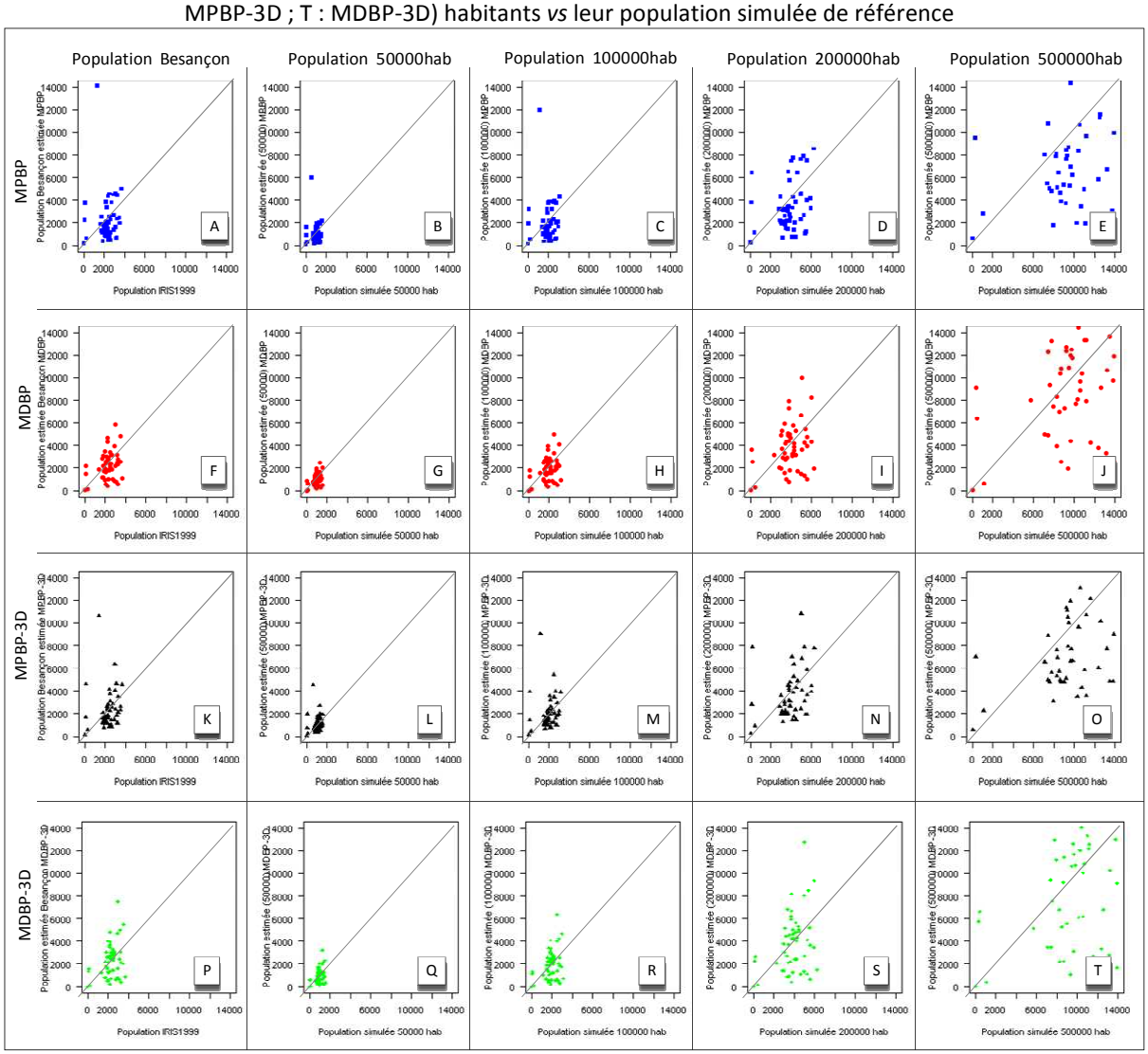
11 G, 11 L, 11 Q ; et fig. 5-12 idem) ; population simulée avec 100 000 habitants (fig. 5-11 C, 11 H, 11 M, 11 R ; et fig. 5-12 idem) ; population simulée avec 200 000 habitants (fig. 5-11 D, 11 I, 11 N, 11 S ; et fig. 5-12 idem) ; et population simulée avec 500 000 habitants (fig. 5-11 E, 11 J, 11 O, 11 T ; et fig. 5-12 idem).

Figure 5-11 : Nuages de points pour la population totale de Besançon en relation avec le recensement de l'INSEE (A : MPBP, F : MDBP ; K : MPBP-3D ; P : MDBP-3D), et pour quatre autres populations simulées avec 50 000 (B : MPBP, G : MDBP ; L : MPBP-3D ; Q : MDBP-3D), 100 :000 (C : MPBP, H : MDBP ; M : MPBP-3D ; R : MDBP-3D), 200 000 (D : MPBP, I : MDBP ; N : MPBP-3D ; S : MDBP-3D), et 500 000 (E : MPBP, J : MDBP ; O : MPBP-3D ; T : MDBP-3D) habitants *vs* le recensement de l'INSEE



En analysant la distribution des nuages de points dans ces graphiques, on observe que dans le premier cas (population selon l'INSEE comme référence, sur l'axe horizontal – fig. 5-11) la forme du nuage des points se maintient, mais un étalonnage sur l'axe vertical est visible, proportionnel à la taille de la population simulée. Ce comportement se répète quelle que soit la méthode employée pour estimer les populations.

Figure 5-12 : Nuages de points pour la population totale de Besançon en relation avec le recensement de l'INSEE (A : MPBP, F : MDBP ; K : MPBP-3D ; P : MDBP-3D), et pour quatre autres populations simulées avec 50 000 (B : MPBP, G : MDBP ; L : MPBP-3D ; Q : MDBP-3D), 100 :000 (C : MPBP, H : MDBP ; M : MPBP-3D ; R : MDBP-3D), 200 000 (D : MPBP, I : MDBP ; N : MPBP-3D ; S : MDBP-3D), et 500 000 (E : MPBP, J : MDBP ; O : MPBP-3D ; T : MDBP-3D) habitants *vs* leur population simulée de référence



Pour le second cas (population de référence simulée, sur l'axe horizontal – fig. 5-12), la forme du nuage des points se maintient, et cette fois-ci l'étalonnage est proportionnel sur les deux axes (horizontal et vertical). De même, ce comportement est répétitif, peu importe la méthode employée pour estimer les populations.

De cette analyse, on peut déduire que la corrélation des populations estimées avec celle de référence se maintient en dépit de la modification de la taille de la population. En effet, cette variation de taille se traduit seulement par un étalonnage, qui peut être interprété comme un changement d'échelle dans le nuage de points. Ainsi, la population totale de la ville étant un paramètre global requis dans notre modèle de population, ce paramètre peut être approximatif car la distribution de la population par la méthode

dasymétrique est indépendante de lui, même s'il sert à marquer l'échelle. En conséquence, la distribution de la population ne dépend que de la variable employée pour la désagrégation, à savoir la variable « bâti / non bâti » dans notre cas. D'ailleurs, faisons une analogie avec le processus de triangulation aérienne, en cartographie (voir photogrammétrie) : la variable sur laquelle se réalise la désagrégation, permettrait l'exécution de l'orientation relative ; tandis que la population totale permettrait l'exécution de l'orientation absolue.

Une deuxième expérience a été réalisée avec les populations simulées, utilisant seulement celles estimées par la méthode du pixel du bâti pondéré en 3D, révélée la plus performante, d'après toutes nos analyses. Les taux bruts d'incidence ont été calculés et représentés -à travers des cartes de distribution des taux bruts d'incidence- tant pour le cancer du sein (fig. 5-13 de A à E) que pour le lymphome non-hodgkinien (fig. 5-13 de F à J). Pour ce faire, les cas d'incidence restent inchangés dans toute la simulation, pendant que les populations estimées varient. Dans les cartes, la distribution des taux bruts d'incidence a été symbolisée par quartile.

Dans les graphiques, nous observons que la distribution spatiale des taux bruts d'incidence ne change pas : autrement dit, les IRIS qui constituent chacun des quatre quartiles sont les mêmes dans chacune des populations simulées. Certes, les valeurs des taux varient, de façon inversement proportionnelle à la taille de la population totale. Par exemple, les taux bruts d'incidence haute -dernier quartile- pour le cancer du sein (chez les femmes) sont les suivants : pour la population réelle de Besançon (à savoir 117 661 habitants), c'est entre 114 et 216 (fig. 5-13 A) ; pour une population simulée de 50 000 habitants, c'est entre 267 et 508 (fig. 5-13 B) ; pour une population simulée de 100 000 habitants, c'est entre 134 et 254 (fig. 5-13 C) ; pour une population simulée de 200 000 habitants, c'est entre 67 et 127 (fig. 5-13 D) ; et pour une population simulée de 500 000 habitants, c'est entre 28 et 51 (fig. 5-13 E). Ce comportement se répète dans les trois autres quartiles : c'est encore vérifié pour la distribution spatiale du lymphome non-hodgkinien.

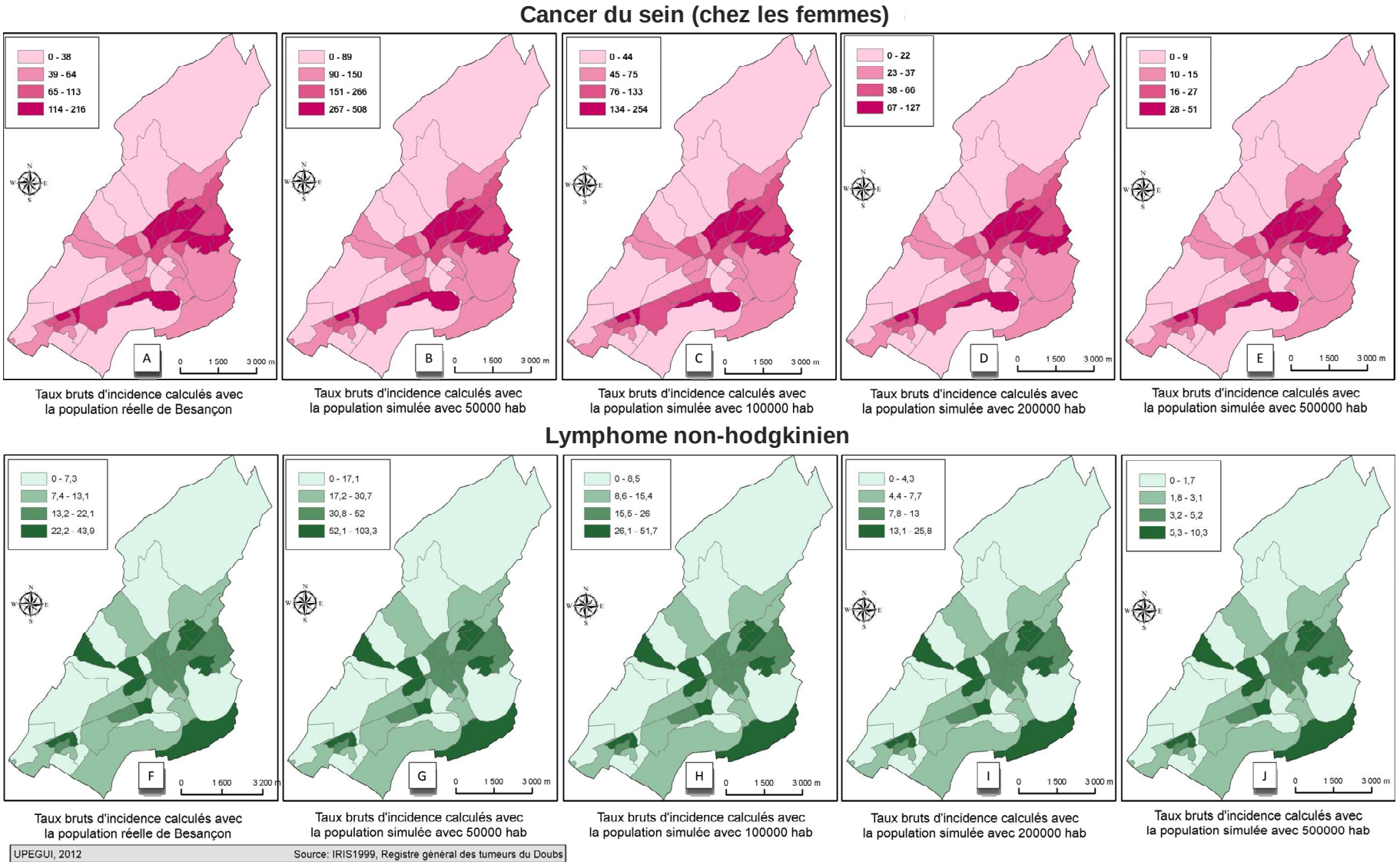
Ces résultats n'ont rien d'étonnant quand on considère la proportionnalité des équations de la méthode dasymétrique quant à la population totale. On peut donc conclure, d'une part, que le paramètre « population totale de la ville » qui *a priori* pourrait être une

limite pour la reproductibilité de notre méthode, est secondaire car son rôle est de fixer les limites d'échelonnage de la population.

D'autre part, l'identification et la distribution spatiale des zones avec un taux brut d'incidence élevé (ou faible), dans une région donnée et avec un nombre de nouveaux cas donné, ne change pas, ou seulement de façon relative, quand la taille de la population change. Ainsi, dans le cas où le recensement n'est pas connu (ou connu mais non fiable), l'épidémiologiste peut tout de même se servir de notre méthodologie : les résultats des taux bruts d'incidence estimés seront alors relatifs. Nous rejoignons ainsi les résultats obtenus dans la recherche de Tran et de son équipe (2002), avec leur cartographie des taux d'incidence relatifs d'une maladie donnée (la fièvre Q dans les environs de Cayenne, en Guyane française) en utilisant les données de télédétection. En revanche, si les effectifs de la population sont connus et fiables, mais que l'échelle géographique des unités spatiales à laquelle ces effectifs se réfèrent n'est pas détaillée, l'épidémiologiste peut se servir de notre méthodologie pour avoir des résultats à un niveau infra-communal. Les résultats des taux bruts d'incidence estimés seront absolus : nous convergeons également avec les résultats de la recherche menée par Viel et Tran (2009). En substance, dans les deux cas, les zones de grande incidence seront identifiées -soit de manière relative, soit de manière absolue-, et les décideurs pourront ainsi adapter l'offre de soins aux besoins de santé de la communauté étudiée.

Par ailleurs, les différentes analyses de sensibilité mettent en évidence la facilité d'adaptation de la méthode proposée, répondant ainsi aux attentes du point de vue de la santé publique (comme déjà dit : simple, automatique, reproductible et exportable).

Figure 5-13. Cartes de distribution du taux brut d'incidence (par 100 000 hab.) calculé : pour les cancers du sein de A à E, et pour les lymphomes non hodgkiniens de F à J ; et ce utilisant: la population réelle de Besançon (A et F) ; et les quatre populations simulées avec 50 000(B et G) ; 100 000 (C et H) ; 200 000(D et I) ; et 500 000(E et J) hab.



Nous avons présenté dans ce chapitre le calcul des taux bruts d'incidence par utilisation des populations estimées (voir chapitre 4) à partir de la variable « bâti / non bâti » -issue des traitements télédétection sur des données à THRS. Ces calculs ont été réalisés pour deux maladies différentes : le cancer du sein chez les femmes, maladie plus fréquente ; et le lymphome non hodgkinien, maladie rare. Ainsi, pour chacune de ces deux maladies, 21 taux bruts d'incidence différents ont-ils été calculés : le premier calculé à partir de la population selon l'INSEE 1999, utilisé comme taux de référence ; et les 20 autres calculés à partir des populations estimées. Les résultats obtenus lors de ce calcul ont été validés avec les taux de référence à travers le CCI. De plus, ils ont été représentés par des graphiques bi-variés (taux observés *versus* taux estimés). Des analyses de sensibilité ont été conduites afin d'établir les limites de notre méthode. Ces analyses ont mis en relief, entre autres : l'importance de disposer d'effectifs de cancers suffisants pour conduire à des taux plus précis lors de l'utilisation de notre méthode ; et le faible rôle de la population totale -paramètre global nécessaire pour la modélisation dasymétrique de la population- sur l'estimation des taux d'incidence.

Conclusion

L'objectif principal de cette thèse était de mettre au point une méthodologie que les épidémiologistes pourraient utiliser dans les cas où les effectifs de population seraient indisponibles, ou peu fiables, ou insuffisamment détaillés pour un usage épidémiologique. On sait que « l'épidémiologie est considérée comme une science des dénominateurs, parce qu'une connaissance précise de la population à risque est une condition fondamentale pour déterminer et dériver des indicateurs significatifs de l'état de santé, des services de santé ainsi que du système de santé » (Viel et Tran, 2008). Cette méthodologie visait donc à estimer la population à partir des données télédétection à THRS, pour permettre l'estimation des indicateurs de santé d'une communauté, en particulier des taux bruts d'incidence. Certes, comme nous l'avons exposé, la problématique de l'estimation de la population a été étudiée depuis longtemps, mais néanmoins, dans cette recherche apparaissent deux aspects novateurs : d'une part, l'utilisation des données à THRS ; d'autre part, l'application à la santé publique. La mise à disposition pour une utilisation civile, en 1999, d'images possédant un mètre de résolution spatiale, a suscité un regain d'intérêt par la communauté scientifique confronté à l'estimation de populations (entre autres). De plus, des améliorations (au niveau de la résolution spatiale, spectrale, radiométrique et temporale) ainsi que de nouvelles technologies n'ont cessé d'être proposées dans ces dernières décennies. L'épidémiologie, quant à elle, s'est servie depuis une vingtaine d'années de la télédétection : dans un premier temps, pour l'étude des maladies infectieuses (plutôt en zone intertropicale) ; plus récemment, pour la mesure d'expositions environnementales

(pesticides, vague de chaleur) ; et dernièrement, pour l'estimation de populations humaines.

C'est dans ce dernier cadre que s'est insérée notre recherche, qui présente ainsi une dimension transdisciplinaire (télédétection, épidémiologie et analyse spatiale), dans la continuité de Tran et de son équipe (2002) qui ont proposé le calcul des taux d'incidents bruts relatifs, et de Viel et Tran (2009), qui ont développé le calcul de taux d'incidence bruts absolus. Compte tenu de cette transdisciplinarité, nous avons également tiré parti de travaux en provenance d'autres champs disciplinaires. Pour l'analyse spatiale, la recherche de Wu et son équipe (2008), nous a permis de cibler la méthode dasymétrique afin de modéliser la population. Pour la télédétection, les travaux Benediktsson et ses collaborateurs (2003) ont été notre clé pour l'extraction par morphologie mathématique ; ceux d'Ackermann et Mering (2007), nous ont permis de mettre au point l'approche de la classification orientée-objets non dirigée, notamment, celle ascendante hiérarchique; ainsi ceux de Dong et de son équipe (2011), nous ont inspiré pour la classification dirigée fondée sur les pixels (plus précisément, l'approche hiérarchisée).

Notre méthodologie est fondée sur l'hypothèse d'un lien existant entre la densité urbaine et l'organisation spatiale du bâti: elle consiste en trois grandes étapes. La première étape a consisté en l'extraction des bâtiments à partir des données télédétection THRS ; la deuxième en l'estimation de la population par modélisation dasymétrique ; et la dernière en le calcul des taux bruts d'incidence, avec comme dénominateur les populations estimées à partir des bâtiments extraits (nous y reviendrons plus loin). Chacune de ces étapes a disposé d'une « vérité terrain », permettant la validation des résultats obtenus et la mise au point de la méthodologie. De plus, dans chacune d'elles, différentes solutions ont été testées avec différents types de données, afin d'appréhender un large éventail de scénarios possibles. En outre, en tant que test méthodologique, nous avons eu recours à l'habitat bisontin, avec sa diversité dans les densités de constructions, les types d'habitat et la couverture du sol. Cette diversité offre une gamme de choix intéressante du point de vue de l'exportabilité.

Pour la première étape (chapitre 3), l'objectif était d'extraire des bâtiments à partir des données THRS. Nous avons utilisé quatre méthodes différentes : en ce qui concerne la classification fondée sur le pixel, l'algorithme ISODATA et la classification hiérarchisée ont été employés ; pour la classification fondée sur l'objet, la classification ascendante hiérarchique et la régression logistique ont été utilisées. Avec cette expérimentation, nous avons mis en relief le rôle important que joue la hauteur dans l'identification des surfaces bâties. Ainsi, la classification hiérarchisée s'est révélée la plus rapide, la plus simple à mettre en place, la plus compréhensible par l'utilisateur, et enfin la plus performante. Néanmoins, en termes d'exportabilité, cette méthode a des limites dues à son coût. C'est pour cette raison que d'autres méthodes ont été proposées comme alternatives afin que l'utilisateur de notre méthodologie puisse s'y référer s'il dispose de ressources limitées. Les résultats obtenus avec les trois autres classifications (ISODATA, CAH et RL) ont été similaires quant à la précision globale et quant au coefficient kappa. Cependant, la classification ISODATA a surestimé les bâtiments car elle inclut : les routes, les sols nus, autrement dit, les zones imperméables ; en revanche, les classifications CAH et par RL ont sous-estimé les bâtiments. En effet, ces deux classifications sont fondées sur la morphologie mathématique et sur les attributs morphologiques des objets. De plus, la diversité que représente la ville rend difficile une extraction indifférenciée des typologies des bâtiments. En définitive, la classification fondée sur l'objet (CAH et RL) a permis de s'approcher de la forme réelle des bâtiments de manière plus précise qu'avec ISODATA. En outre, la classification par RL est un choix plus déterministe que celles d'ISODATA et de la CAH, requérant plus de décisions de la part de l'opérateur. Par ailleurs, le temps de calcul et la place pour stocker les résultats intermédiaires de l'approche orientée-objet ont été bien supérieurs à ceux demandés avec ISODATA.

En ce qui concerne la modélisation de la population par la méthode dasymétrique (chapitre 4), notre objectif était double : d'un côté, établir l'influence de la variable « bâti / non bâti » (produite dans la étape précédente) dans la modélisation ; de l'autre, identifier la méthode la plus pertinente pour conduire cette modélisation. Pour ce faire, nous avons utilisé l'ilôt et l'IRIS comme échelles de l'unité géographique « d'arrivée ». Cela nous permet de conclure que le détail de ces unités géographiques est fortement lié au niveau de détail de la variable explicative. C'est-à-dire que, si on part d'une image avec

beaucoup de détail (par exemple 50 cm de résolution spatiale) et qu'on souhaite arriver à une unité géographique moins détaillée (par exemple l'IRIS), on obtiendra le même degré d'association (population estimée *vs* population observée) que si on utilisait une image avec une résolution spatiale moindre (par exemple 30 mètres). Par ailleurs, l'injection de la hauteur -approche volume- permet une différentiation entre les maisons en habitat dispersé et les immeubles, conduisant à la meilleure performance de la MPBP-3D. En outre, dans toutes les analyses réalisées, nous avons observé une variation de la performance de la modélisation en fonction du type de bâtiment, autrement dit, de la classification « bâti / non bâti ». Cela permet de conclure que -malgré les résultats semblables des classifications ISODATA, CAH et RL dans la précision globale comme dans le coefficient kappa- il est nécessaire de réaliser une extraction des bâtiments la plus réaliste possible afin de garantir une modélisation correcte de la population. Sinon, on commet soit l'erreur de placer une population l'a où il n'en existe pas, soit l'erreur inverse de négliger une population là où elle existe. Ainsi, la modélisation de la population fondée sur le bâti extrait par la classification hiérarchisée, en utilisant à la fois la hauteur et le nombre de pixels bâtis comme indicateur de répartition (MPBP-3D) s'est révélée la plus adéquate. Une modélisation de la population s'appuyant sur l'approche surface (en raison des données disponibles) peut néanmoins constituer une alternative raisonnable.

Le calcul des taux d'incidence (chapitre 5) est le point où convergent tous les traitements précédents, dans une démarche transdisciplinaire. Les taux d'incidence de deux maladies, le cancer du sein chez les femmes (maladie relativement fréquente) et le lymphome non-hodgkinien (maladie plutôt rare), ont été calculés en utilisant les différentes populations estimées. Ces taux estimés ont été corrélés avec les « taux de référence », calculés à partir des effectifs de population obtenus par le recensement (INSEE 1999). De manière analogue à l'étape précédente, différentes analyses de sensibilité ont été effectuées. Les résultats obtenus nous ont permis de mettre en évidence un fort degré d'association entre les taux estimés -à partir des populations fondées sur les bâtis extraits par télédétection- et les taux observés. Nous avons également démontré que la surestimation de la population due à l'utilisation non résidentielle des sols n'avait pas d'influence sur les calculs des taux d'incidence. Par ailleurs, l'effet de la fluctuation aléatoire liée aux petits effectifs a été mis en évidence

pour le lymphome non-hodgkinien. Cette fluctuation est accrue par l'incertitude sur la présence de bâtiments, qui trouve son origine, d'une part, dans la surestimation des bâtis par ISODATA ; et de l'autre part, dans la sous-estimation des bâtis par CAH et par RL. C'est pourquoi nous recommandons l'application de notre méthodologie à une maladie suffisamment fréquente. Enfin, nous avons conclu que la connaissance de la population totale, pré-requise dans la méthode dasymétrique, peut n'être approximative au prix d'une estimation de taux relatifs et non plus absolus.

Nous avons prouvé à travers les valeurs élevées du CCI (taux estimés *vs* taux observés), que les données télédétection couplées à la modélisation de la population (analyse spatiale) permettent le calcul de taux bruts d'incidence. Les données altimétriques ont joué un rôle très important tant dans l'extraction des bâtiments que dans l'estimation des populations et par suite dans l'estimation de ces taux. Cependant, la méthodologie est aussi utilisable dans des zones ne possédant pas de données altimétriques, certes avec un degré d'association moindre mais encore déterminant. Ces résultats soulignent le potentiel de la télédétection pour estimer des populations humaines, et de là, l'estimation d'indicateurs de santé (Tran *et al.*, 2002 ; Tran *et al.*, 2004 ; Viel et Tran, 2009). Notre méthode apparaît transposable dans d'autres zones d'étude, proposant différentes alternatives en fonction de la disponibilité des données.

Du point de vue de la santé publique, différentes perspectives sont ouvertes, en particulier l'estimation d'une population à risque environnemental, ou la détection rapide de zones présentant des taux d'incidence élevés. Ainsi, dans le cadre d'une catastrophe environnementale (rejets toxiques dans l'atmosphère, par exemple), la population exposée (c'est-à-dire localisée sous le panache) pourrait être estimée par notre méthode, en réagrégeant la population selon le polygone du panache, sous réserve d'un accès rapide aux données télédétection (prévue dans la Charte internationale pour la gestion des catastrophes). Dans le cas d'une épidémie, notre méthode permettrait d'estimer son extension et d'identifier les zones les plus affectées par une maladie infectieuse transmissible, permettant aux autorités sanitaires et aux associations humanitaires d'intervenir plus précocement.

In fine, notre méthode permet l'identification des zones d'incidence élevée, fondées sur les populations estimées à partir de données télédétection, répondant ainsi aux besoins de données démographiques de la part des épidémiologistes, dans l'éventualité d'une absence de données plus précises. Elle se révèle constituer un outil innovant pour les épidémiologistes, leur permettant de mesurer l'état de santé de la population à une échelle fine. Les taux d'incidence estimés, transmis aux décideurs de santé publique, constituent des éléments d'aide à la décision pour la planification sanitaire, permettant d'adapter l'offre de soins aux besoins de santé.

Bibliographie

A

ACKERMANN G., MERING C., 2007, « Measuring morphology of urban settlements from high resolution remote sensing images », *XIIth Intern. Congress of Stereology*, Saint Etienne (France), 4-7 septembre 2007, http://icsxii.univ-st-etienne.fr/

AKAIKE, H., 1974. « A new look at statistical model identification », *IEEE Transactions on Automatic Control AU-19*, pp. 716-722.

ANDERSON D., ANDERSON P., 1973, « Population estimation by Humans and machines », *Photogrammetric Engineering*, 39, pp 147-154.

ANDERSON J., 1971, « Land-Use Classification Schemes -used en selected recent geographic applications of remote sensing », *Photogrammetric Engineering*, vol. 37, pp. 379-387.

B

BAJAC Q., 2010, La *photographie du daguerréotype au numérique*, Editions Gallimard, France : 383 p.

BAKIS H., 1978, *La photographie aérienne et spatiale, Que sais-je ?*, Presses Universitaires de France, Paris : 128 p.

BALL G., HALL D., 1965, « ISODATA, A Novel Method of Data Analysis and Pattern Classification », *Stanford Research Institute*, Menlo Park, California : 61 p.

BARRETT F., 2000, « Finke's 1792 map of human diseases: the first world disease map? » *Social Science & Medicine*, vol. 50, pp. 915-921.

BECK L., LOBITZ B., WOOD B., 2000, « Remote Sensing and Human Health: New Sensors and New Opportunities », *Emerging Infectious Diseases*, vol. 6, pp. 217-227.

BENEDIKTSSON J., PESARESI M., ARNASON K., 2003, « Classification and feature extraction for remote sensing images from urban areas based on morphological transformations », *IEEE Transactions on Geoscience and Remote Sensing*, vol. 41, n° 9, pp. 1940-1949.

BENZECRI J.P. & collaborateurs, 1980, *L'analyse des données : 1 La Taxonomie*, DUNOD, Paris : 625 p.

BERNARD P-M., LAPOINTE C., 1987, *Mesures statistiques en épidémiologie*, Presses de l'Université du Québec, Québec : 314 p.

BERKSON J., 1944, « Application of the Logistic Function to Bio-Assay », *Journal of the American Statistical Association*, vol. 39, No. 227, pp. 357-365

BERTIN J., 2005, Sémiologie *graphique les diagrammes -les réseaux -les cartes*, Les réimpressions des Editions de l'Ecole des Hautes Etudes en Sciences Sociales, 4ème édition, Paris : 452 p.

BEUCHER S., 1990, *Segmentation d'images et morphologie mathématique,* Thèse de doctorat, Ecole nationale Supérieure des Mines de Paris, Paris, 294 p.

BEUCHER S., LANTUEJOUL., 1979, « Use of watersheds in contour detection », Int. Workshop on Image Processing, CCETT/IRISA, Rennes, France, septembre.

BIDALOT G., 2009, *Besançon des origines à nos jours : Histoire politique et économique d'une ville*, Les Presses de Belvédère, France : 206 p.

BIROT M., 1958, « Un recensement de femmes au royaume de Mari », *Syria. Archéologie, Art et histoire*, Tome 35 fascicules 1-2, pp. 9-26.

BLASCHKE T., 2010, « Object based image analysis for remote sensing », *ISPRS Journal of Photogrammetry and Remote Sensing*, n° 65, pp. 2-16.

BLASCHKE T., BURNETT C., PEKKARINEN A. « Image Segmentation Methods for Object-based Analysis and Classification » in DE JONG S., VAN DER MEER F. ("dir"), 2005, *Remote Sensing Image Analysis: Including The Spatial Domain*, Kluwer Academic Publishers, Springer Science, USA, 359 p.

BLOCH I., 2008, *Représentations discrètes et morphologie mathématique, Polycopié de l'UE RDMM*, Université Pierre et Marie Curie, ENST, 53 p.
[en ligne, consulté le 25 mars 2012 http://www.tsi.enst.fr/~bloch/P6Image/rdmm.pdf]

BONNE J., MCGWIRE K., OTTESON E., DEBACA R., KUHN E., VILLARD P., BRUSSARD P., JEOR S., 2000, « Remote sensing and geographic information systems : Charting sin nombre virus infections in Deer Mice », *Emerging Infectious Diseases*, vol. 6, pp. 248-258.

BOUYER J., HEMON D., CORDIER S., DERRIENNIC F., STÜCKER I., STENGEL B., CLAVEL J., 1995, *Épidémiologie principes et méthodes quantitatives*, Les Editions INSERM, Paris : 498 p.

BRACKEN I., 1991, « A Surface model approach to small area population estimation », *The Town Planning Review*, 62, n° 2, pp. 225-237.

BRACKEN I., MARTIN D., 1989, « The generation of spatial population distributions from census centroïd data », *Environment and Planning A,* 21, n°4, pp. 537-543.

BRIGGS D., GULLIVER J., FECHT D., VIENNEAU D., 2007, « Dasymetric modelling of small-area population distribution using land cover and light emissions data », *Remote Sensing of Environment, 108*, pp. 451-466.

BRUNET R., 1980, « La composition des modèles dans l'analyse spatiale », *Espace géographique*, vol 9, n° 4, pp 253-265.

C

CALVAT G., 2003, *La maison de A à Z : Le vocabulaire de la construction*, Editions alternatives : 176 p.

CARBONNELL M., 1968, *Panorama des applications de la photographie aérienne*, Mémoires de photo-interprétation V, Ecole Pratique des Hautes Etudes -VIe section centre de recherches historiques, S.E.V.P.E.N, Paris : 57 p.

CENTERS FOR DISEASE CONTROL AND PREVENTION (CDC), 2006, Principles of epidemiology in Public Health Practice, an introduction to applied epidemiology and biostatistics, Course SS1000, Third Editions, US,
[en ligne, consulté le 25 mars 2011, http://www.cdc.gov/training/products/ss1000/ss1000-ol.pdf]

CHAMOUTON C., 2009, *Les maisons comtoises*, Collection Reflets de Terroir, Editions CPE, France : 128 p.

CHAMPELY S., 2006, *Tests statistiques paramétriques : Puissance, taille d'effet et taille d'échantillon (sous R), Polycopié du cours*, Université Lyon 1, France : 43 p.
[en ligne, consulté le 24 mai 2012 http://pbil.univ-lyon1.fr/R/pdf/puissance.pdf]

CHANUSSOT J., BENEDIKTSSON J., FAUVEL M., 2006, « I Classification of Remote Sensing Images From Urban Areas Using a Fuzzy Possibilistic Model », IEEE *Geoscience and Remote Sensing Letters*, vol. 3, n° 1, pp. 40-44.

CHEN Y., SU W., LI J., SUN Z., 2009, « Hierarchical object oriented classification using very high resolution imagery and LIDAR data over urban areas », *Advances in Space Research*, n° 43, pp. 1101-1110.

CHESNAIS J-C., 1990, *La démographie, Que sais-je ?*, Presses Universitaires de France, Paris : 127 p

CHESSEL D., DUFOUR A., THIOULOUSE J., 2004, « The ade4 package - I : One-table methods », *R News*, vol. 4/1, pp. 5-10.

CHESSEL D., THIOULOUSE J., DUFOUR A., 2004b, « Introduction à la classification hiérarchique », *Fiche de Biostatistique -Stage 7* [en ligne, consulté le 17 avril 2012 http://pbil.univ-lyon1.fr/R/stage/stage7.pdf]

CHEVALLIER, R., 1971, *La photographie aérienne*, Collection U2, Librairie Armand Colin, Paris : 233 p.

CHOPIN F., MERING C., 2004, « Cartographie de la densité du bâti par analyse granulométrique des images de télédétection », *Revue Française de Photogrammétrie et de Télédétection*, n° 173/174, pp. 113-122.

CHUVIECO E., 1996, *Fundamentos de Teledetección espacial*, 3ra Edición. Ed. Rialp, Madrid : 568 p.

CLINE B., 1970, « New eyes for epidemiologists: aerial photography and other remote sensing techniques ». *American Journal of Epidemiology*, vol. 92, pp. 85-89.

COHEN J., 1960, « A coefficient of agreement for nominal scales », *Educational and Psychological Measurement*, vol. 20, pp. 27-46.

COMMISSION DE COMMUNAUTÉS EUROPÉENNES, 1993, « CORINE Land Cover technical guide », European Union, (Luxembourg, Directorate-General for the Environment, Nuclear Safety and Civil Protection).

CONSEIL INTERNATIONAL DE LA LANGUE FRANÇAISE (CILF), 1997, *Terminologie de télédétection et photogrammétrie (Français-Anglais)*, PUF, Paris : 455 p.

COSTER M., CHERMANT J-L., 1989, *Précis d'analyse d'images*, Presses du CNRS, Paris : 521 p.

COURTIEU J., 1982, *Dictionnaire des communes du département du Doubs : Tome 1 Abbans-dessous - Bouverans*, Editions CÊTRE, Besançon : 495 p.

CUFF D., MATTSON M., 1982, *Thematic maps: Their design and production*, Methuen & Co, USA : 171 p.

CURRAN P.J., ATKINSON P.M., FOODY G.M., MILTON E.J., 2000, « Linking remote sensing, land cover and disease » *Advances in Parasitology*, vol. 47, pp. 37-80.

CZERNICHOW P., CHAPERON J., LE COUTOUR X., 2001, *Épidémiologie : Connaissances et Pratiques*. Les Editions Masson, Paris : 443 p.

D

DE HUTOROWICZ H., ADLER B., 1911, « Maps of Primitive Peoples », *Bulletin of the American Geographical Society*, vol. 43, No. 9, pp. 669-679.

DE JONG Steven, VAN DER MEER Freek, 2005, *Remote Sensing Image Analysis: Including The Spatial Domain*, Kluwer Academic Publishers, Springer Science, USA : 359 p.

DE LA ROCQUE S., MICHEL V., PLAZANET D., PIN R., 2004, « Remote sensing and epidemiology : examples of applications for two vector-borne diseases », *Comparative Immunology, Microbiology & Infectious Diseases*, 27, pp. 331-441.

DENIS E., MORICONI-EBRARD F., 2009 « La croissance urbaine en Afrique de l'ouest : De l'explosion à la prolifération », *La Chronique du CEPED*, n° 57.

DIRECTION URBANISME ET HABITAT (D.U.H.), 2007, *Plan Local d'Urbanisme : 3.1 Diagnostic*, Ville de Besançon : 168 p.

DOBSON J., BRIGHT E., COLEMAN P., DURFEE R., WORLEY B., 2000, « LandScan : A global population database for estimating populations at risk », *Photogrammetric Engineering & Remote Sensing*, 66, n° 7, pp. 849-857.

DONG P., RAMESH S., NEPALIE A., 2010, « Evaluation of small-area population estimation using LiDAR, Landsat TM and parcel data », *International Journal of Remote Sensing*, 31, n° 21, pp. 5571-5586.

DUMONT G-F., 2004, *Les populations du monde*, Collection U -Géographie -Armand Colin, Paris : 288 p.

DURAND J-M., 2000, « Société du royaume de Mari (XIXème et XVIIIème siècle avant notre ère) », *Résumé de cours année 1999-2000* au Collège de France.
[en ligne, consulté le 25 mars 2011 http://www.college-de-france.fr/media/assyrio/UPL52961_DurandR99.pdf]

E

EICHER C., BREWER C., 2001, « Dasymetric mapping and areal interpolation: Implementation and evaluation », *Cartography and Geographic Information Science*, vol. 28, n°. 2, pp 125-138.

ENCYCLOPEDIA OF BIOSTATISTICS, Volume 4 MED-PRE, 1998, “editors”, Armitage P., Colton T., John Wiley & Son, England.

F

FARR, T.G., KOBRICK M., 2000, « Shuttle Radar Topography Mission produces a wealth of data”, *Amer. Geophys. Union Eos*, vol. 81, pp. 583-585.

FEDERATION NATIONALE DES CENTRES DE LUTTE CONTRE LE CANCER, 2007, *Comprendre le cancer du sein*, Guide SOR SAVOIR PATIENT, FNCLCC, 1er édition, Paris : 113 p.

FERLAY J., SHIN HR., BRAY F., FORMAN D., MATHERS C., PARKIN DM., 2010, GLOBOCAN 2008 v1.2, *Cancer Incidence and Mortality Worldwide*: IARC CancerBase No. 10 [Internet]. Lyon, France: International Agency for Research on Cancer; Available from: http://globocan.iarc.fr, accessed on 11/mai/2010.

FISHER P., LANDFORD M., 1995, « Modelling the errors in areal interpolation between zonal systems by Monte Carlo simulation », *Environment and Planning A*, vol. 27, pp. 211-224.

FLORET N., MAUNY F., CHALLIER B., ARVEUX P., CAHN J-Y., VIEL J-F., 2003, « Dioxin emissions from a solid waste incinerator and risk of non-Hodgkin lymphoma », *Epidemiology*, vol. 14, pp. 392-398.

FLOWERDEW R., GREEN M., 1991, « Data integration : Statistical methods for transferring data between zonal systems », *Handling geographical information : Methodology and potential applications*, ed. I. Masser and M. Blakemore, NewYork, Wiley, pp. 38-54.

FORD T., COLWELL R., ROSE J., MORSE S., ROGERS D., YATES T., 2009, « Using satellite images of environmental changes to predict infectious disease outbreaks », *Emerging Infectious Diseases*, vol. 15, pp. 134-1346.

FORESTIER G., PUISSANT A., WEMMERT C., GANÇARSKI P., 2012, « Knowledge-based region labeling for remote sensing image interpretation », *Environment and Urban Systems,* In Press, doi:10.1016/j.compenvurbsys.2012.01.003

FOROUZANFAR M., FOREMAN K., DELOSSANTOS A., LOZANO R., LOPEZ A., MURRAY C., NAGHAVI M., 2011, « Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis », *The Lancet oncology*, vol. 378, pp.1461-1484.

FOUCART T., 1997, *L'analyse des données : Mode d'emploi*, DIDACT Statistique, Presses Universitaires de Rennes, Rennes, 188 p.

FRANCHE-COMTE - DELEGATION REGIONALE A L'ARCHITECTURE ET A L'ENVIRONNEMENT (DRAE), 1979, *La politique régionale d'amélioration des matériaux de couverture en Franche-Comté*, DRAE, Besançon : 51 p.

FRIENDLY M., DENIS D.J., 2001, « Milestones in the history of thematic cartography, statistical graphics, and data visualization ». [en ligne, consulté le 28 mars 2011 http://www.math.yorku.ca/SCS/Gallery/milestone/]

G

GAGNON H., 1974, *La photo aérienne son interprétation dans les études de l'environnement et de l'aménagement du territoire*, Les éditions HRW Ltée : 278 p.

GAMBA P., DU P., JUERGENS C., MAKTAV D., 2011, « Foreword to the Special Issue on "Human Settlements: A Global Remote Sensing Challenge" », *IEEE Journal of selected topics in applied earth observations and Remote Sensing*, vol. 4, n° 1, pp. 5-7.

GAYON J., 2000, « De la croissance relative à l'allométrie (1918-1936) / From relative growth to allometry (1918-1936) », *Revue d'histoire des sciences*, Tome 53 n° 3-4, pp. 475-498.

GIRARD M.C., GIRARD C.M., 1989, *Télédétection appliquée : zones tempérées et intertropicales*, Collection sciences agronomiques, Masson, Paris : 260 p.

_____, 1999, *Traitement des données de télédétection*, DUNOD, Paris : 260 p.

GOODCHILD M., LAM N. S-N., 1980, « Areal interpolation : A variant of the traditional spatial problem », *Geoprocessing*, 1, pp. 291-312.

GOODCHILD M., ANSELIN L., DEICHMANN U., 1993, « A framework for the areal interpolation of socioeconomic data », *Environment and Planning A*, 25, pp. 383-397.

GRAUNT J., 1665, *Natural and political observations mentioned in a following index, and made upon the Bills of Mortality*, London.

GREEN N., 1956, « Aerial Photographic Analysis of Residential Neighborhoods: An Evaluation of Data Accuracy », *Social Forces,* 35, n° 2, pp. 142-147.

GUEROIS M., 2003, *Les formes des villes européennes vues du ciel : Une contribution de l'image CORINE à la comparaison morphologique des grandes villes d'Europe occidentale*, Thèse de doctorat, Université Paris I Panthéon - Sorbonne, 258 p.

H

HARRIS R., CHEN Z., 2005, « Giving dimension to point locations: urban density profiling using population surface models », *Computers, Environment and Urban System*, 29, pp. 115-132.

HASTIE T., TIBSHIRANI R., FRIEDMAN J., 2008, *The elements of statistical learning*, Springer-Verlag : 763 p.

HARVEY J.T., 2002, « Estimating census district populations from satellite imagery some approaches en limitations », *International Journal of Remote Sensing*, 23, n° 10, pp. 2071-2095.

_____, 2002b, « Population estimation models based on individual TM pixels », *Photogrammetric Engineering & Remote Sensing*, 68, n° 11, pp. 1181-1192.

HAY S., 1997, « Remote sensing and disease control: past, present and future », *Transactions of the royal society of tropical medicine and hygiene*, vol. 91, pp. 105-106.

HERBRETEAU V., SALEM G., SOURIS M., HUGOT J-P., GONZALEZ J-P., « Thirty years of use and improvement of remote sensing, applied to epidemiology: From early promises to lasting frustration », *Health & Place*, 2007, vol. 13, pp. 400–403.

HOLT J., LU H., 2011, « Dasymetric mapping for population and socio-demographic data distribution », in *Urban remote sensing monitoring synthesis and modeling in the urban environment*, "dir" YANG X., Wiley-Blackwell, USA: 388 p.

HOUOT H., 1999, *Approche géographique des nuisances sonores urbaines Méthodologie d'aide à la prise en compte des nuisances sonores en aménagement urbaine Application à la ville de Besançon*, Thèse de doctorat, Université de Franche-Comté, Besançon : 329 p.

HSU S-Y., 1971, « Population estimation : a sample study of inter-census applies aerial photos and USGS quads to an area near Atlanta, Georgia », *Photogrammetric Engineering*, 37, pp. 449-545.

HU Z., 2009, « Spatial analysis of MODIS aerosol optical depth, PM2.5, and chronic coronary heart disease », *International Journal of Health Geographics*, 8:27.

HU Z., RAO R., 2009, « Particulate air pollution and chronic ischemic heart disease in the eastern United States: a county level ecological study using satellite aerosol data », *Environmental Health*, 8:26.

HUXLEY J., 1932, « Problems of relative growth », Methuen and Company Limited, London : p 276.

_____, 1950, « Relative growth and form transformation », *Proceedings of the Royal Society of London. Series B*, Biological Sciences, 137, n° 889, pp. 465-469.

I

IISAKA J., HEGEDUS E., 1982, « Population estimation from Landsat imagery », *Remote Sensing of Environment,* 12, pp. 259-272.

IMHOF E., 1982, *Cartographic relief presentation*, De Gruyter : 389 p.

INSTITUT D'AMÉNAGEMENT ET D'URBANISME DE LA RÉGION D'ILE DE FRANCE, 2005, « Note rapide sur l'occupation du sol », n°. 383, Paris : 4°p.

INSTITUT GEOGRAPHIQUE NATIONAL (IGN), 2009, *BD TOPO® Version 2 Descriptif de contenu*, IGN, Paris : 172 p.

INSTITUT NATIONAL DE LA SANTE ET DE LA RECHERCHE MEDICALE (Inserm), 2005, *Cancers Pronostics à long terme* (ouvrage collectif), les éditions Inserm, Paris : 311 p.

_____, 2008, *Cancer et environnement* (ouvrage collectif), les éditions Inserm, Paris : 889 p.

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ETUDES ECONOMIQUES (INSEE), 2005, «Pour comprendre le recensement de la population », Méthodes hors série, 01/05/2005. [en ligne, consulté le 28 avril 2011 http://www.insee.fr/fr/publications-et-services/sommaire.asp?codesage=imeths01]

_____, 2009, « Pourquoi un ajustement est-il parfois introduit dans les variations de population ? »
[en ligne, consulté le 28 avril 2011 http://www.insee.fr/fr/methodes/sources/pdf/Ajustement_et_variations_de_population.pdf]

_____, 2010, Recensement de la population Évolutions : pourquoi privilégier les évolutions par rapport à 1999 ?, [en ligne, consulté le 2 mars 2012 http://www.insee.fr/fr/publics/communication/recensement/particuliers/doc/fiche-evol_2006-%202007.pdf]

J

JACQUIN A., MISAKOVA L., GAY M., 2008, « A hybrid object-based classification approach for mapping urban sprawl in periurban environment », *Landscape and Urban Planning*, n° 84, pp. 152-165.

JAHJAH M., ULIVIERI C., 2010, « Automatic archaeological feature extraction from satellite VHR images », *Acta Astronautica*, n° 66, pp. 1302-1310.

JEMAL A., BRAY F., CENTER M., FERLAY J., WARD E., FORMAN D., « Global Cancer Statistics », *CA: A Cancer Journal for Clinicians*, vol. 61, n°.2, pp.69–90.

JOLY D., 1990, « Structures cognitive, niveaux d'échelles et statistiques dans l'espace et le temps des géographes », *Geopoint 90*, pp. 179-182.

K

KALLURI S., GILRUTH P., ROGERS D., SZCZUR M., 2007, « Surveillance of Arthropod Vector-Borne Infectious Diseases Using Remote Sensing Techniques: A Review », *PLoS Pathogens*, vol. 3 n° 10, pp. 1361-1371.

KING R., CAMPBELL-LENDRUM D., DAVIES C., 2004, « Predicting geographic variation in cutaneous leishmaniasis, Colombia », *Emerging Infectious Diseases*, vol. 10, n°4, pp. 598-607

KRAUS S., SENGER L., RYERSON J., 1974, « Estimating population from photographically determinated residential land use types », *Remote Sensing of Environment*, 3, pp. 35-42.

L

LAAIDI K., ZEGHNOUN A., DOUSSET B., BRETIN P., VANDENTORREN S., GIRAUDET E., BEAUDEAU P., 2012, « The Impact of Heat Islands on Mortality in Paris during the August 2003 Heat Wave », *Environmental Health Perspectives*, vol. 120, n°2, pp. 254-259.

LALANNE L., 1845, « Remarques à l'occasion du Mémoire de M. Morlet sur les centres de figures ; et réflexions sur la représentation graphique de divers éléments relatifs à la population », *Comptes rendus des Séances de l'Académie des Sciences*, 20, pp. 438-441.

LAM N. S-N, 1983, « Spatial interpolation methods: A review », *The American Cartographer,* 10, n° 2, pp. 129-149.

LANDIS J., KOCH G., 1977, « The Measurement of Observer Agreement for Categorical Data », *Biometrics*, vol. 33, n° 1, pp. 159-174.

LANGFORD M., 2006, « Obtaining population estimates in non-census reporting zones : An evaluation of the 3-class dasymetric method », *Computers, Environment and Urban System*, 30, pp. 161-180.

_____, 2007, « Rapid facilitation of dasymetric-based population interpolation by means of raster pixel maps », *Computers, Environment and Urban Systems*, vol. 31, pp. 19-32.

LANGFORD M., HIGGS G., RADCLIFFE J., WHITE S., 2008, « Urban population distribution models and service accessibility estimation », *Computers, Environment and Urban System*, 32, pp. 66-80.

LAURENCELLE L., 2007, « Inventer ou estimer la puissance statistique ? Quelques considérations utiles pour le chercheur », *Tutorials in Quantitative Methods for Psychology*, Vol. 3(2), pp. 35-42.

LEE D., SALLEE G., 1970, « A Method of Measuring Shape », *Geographical Review*, vol. 60, n° 4, pp. 555-563.

LHOMME S., 2005, *Identification du bâti à partir d'images satellitaires à très hautes résolutions*, Thèse de doctorat Université Louis Pasteur I, Strasbourg : 258 p.

LI G., WENG Q., 2005, « Using Landsat ETM+ Imagery to Measure Population Density in Indianapolis, Indiana, USA », *Photogrammetric Engineering & Remote Sensing*, 71, n° 8, pp. 947-958.

LINDGREN D., 1971, « Dwelling unit estimation with color-IR photos », *Photogrammetric Engineering*, 37, pp. 373-377.

LIU X., CLARKE K., HEROLD M., 2006, « Population Density and Image Texture : A comparison study », *Photogrammetric Engineering & Remote Sensing*, 72, n° 2, pp. 187-196.

LO C.P., WELCH R., 1977, « Chinese Urban Population Estimates », *Annals of the Association of American Geographers*, 67, n° 2, pp. 246-253.

LO C.P., 1989, « A raster approach to population estimation using high-altitude aerial and space photographs », *Remote Sensing of Environment,* 27, pp. 59-71.

_____, 1992, « A GIS approach to population estimation in a complex urban environment using SPOT multispectral images », XVII Congress International Society for Photogrammetry and Remote Sensing, Volume XXIX, Part B7, Commission VII, Washington, USA, pp. 935-941.

_____, 1997, « Application of Landsat TM data for quality of life assessment in an urban environment » *Compt, Environ and Urban System*, vol. 21, pp. 259-276.

_____, 2001, « Modeling the population of China using DMSP Operational Linescan System nighttime data », *Photogrammetric Engineering & Remote Sensing*, 67, n° 9, pp. 1037-1047.

LONGLEY P., GOODCHILD D., MAGUIRE D., RHIND D., 2011, *Geographic Information Systems and Science,* John Wiley & Sons, USA : 539 p.

LU Z., IM J., QUACKENBUSH L., HALLIGAN K., 2010, « Population estimation based on multi-sensor data fusion », *International Journal of Remote Sensing*, 31, n° 21, pp. 5587-5604.

LWIN K., MURAYAMA Y., 2011, « Estimation of building population from LiDAR derived digital volume model » in *Spatial analysis and modeling in geographical transformation process: GIS-based application*, "dir" MURAYAMA Y., THAPA R., Springer, New York : 300 p.

M

MAANTAY J., MAROKO A., HERRMANN Christopher, 2007, « Mapping population distribution in the urban environment : The cadastral-based expert dasymetric system (CEDS) », *Cartography and Geographic Information Science*, 34, n° 2, pp. 77-102.

MACEACHREN A., 1979, « The evolution of thematic cartography : A research methodology and historical review », *The Canadian Cartographer*, 16, n° 1, pp. 17-33.

_____, 1994, *Some truth with maps: A primer on symbolization and design*, Association of American Geographers, Washington D.C. : 129 p.

MACIEJEWSKI R, 2011, *Data representations, transformations, and statistic for visual reasoning*, Morgan & Claypool publishers, USA : 86 p.

MANAKOS I., SCHNEIDER T., AMMER U., 2000, « A comparison between the ISODATA and the eCognition classification methods on basis of field data », *XIXth ISPRS Congress, International Archives of Photogrammetry and Remote Sensing. vol. XXXIII, Supplement B7*. Amsterdam, pp. 133-139.

MANGUIN S., BOUSSINESQ M., 1999, « Apport de la télédétection en santé publique : l'exemple du paludisme et autres perspectives », *Med Mal Infect*, Vol 29, pp. 318 -324.

MAROKO A., MAANTAY J., GRADY K., 2011, « Using geovisualization and geospatial analysis to explore respiratory disease and environmental health justice in New York city », pp. 39 -66, in *Geospatial analysis of environmental health*, MAANTAY J.,

MCLAFFERTY S., « Rédacteurs », Series: Geotechnologies and the Environment, Springer, 1st Edition : 498 p.

MARTIN D., 1989, « Mapping population data from zone centroid locations », *Transactions of the Institute of British Geographers, New Series*, 14, n° 1, pp. 90-97.

MATHERON G., 1975, *Random sets and integral geometry*, Wiley and sons, New York : 261 p.

MATSUYAMA T., HWANG, V.-S., 1990, *SIGMA: A Knowledge-Based Aerial Image Understanding System*, PLENUM Press, New York, USA.

MAXWELL S., AIROLA M., NUCKOLS J., 2010, « Using Landsat satellite data to support pesticide exposure assessment in California », *International Journal of Health Geographics*, 9:46.

MELESSE A., JORDAN J., 2002, « A comparison of fuzzy vs. augmented-ISODATA classification algorithms for cloud-shadow discrimination from Landsat Images », *Photogrammetric Engineering & Remote Sensing*, vol. 68, pp. 905–911.

MENNIS J., 2003, « Generating surface models of population using dasymetric mapping », *The Professional Geographer*, vol. 55, n° 1, pp 31-42.

MERING C., 1990, « Présentation de quelques méthodes de la Morphologie Mathématique permettant de caractériser une structure sur une image binaire », *ORSTOM - Journées Télédétection Bondy - Analyse quantitative des formes (SATFORM)*, pp. 193-211.

MERING C., CHOPIN F., 2002, « Granulometric maps from high resolution satellite images », *Image Anal Stereol*, n° 21, pp. 19-24.

MENNIS J., 2003, « Generating surface models of population using dasymetric mapping », *The Professional Geographer*, vol. 55, n° 1, pp. 31-42.

MERRILL R., 2008, *Environmental epidemiology : Principles and methods*, Jones and Bartlett Publishers, Massachusetss : 483 p.

MIELLET P., 1992, « Quelques idées sur le lien entre systèmes d'information géographique et la modélisation en géographie », Géopoint 92 « Modèles et modélisation géographique », Avignon 1992, Actes du colloque, pp 137-139.

MIGNET Ch., 1956, « Le recensement de la Chine - Méthodes et principaux résultats », *Population*, 11e année, n° 4, pp. 725-736. « Traduit par »

MORICONI-EBRARD F., 1991, « Les 100 plus grandes villes du monde », *Economie et statistique*, n° 245, pp. 7-18.

N

NAGAO M., MATSUYAMA T., IKEDA Y., 1979, « Region extraction and shape analysis in aerial photographs », *Computer Graphics and Image Processing*, vol.10, n° 3, pp. 195-223.

NAGAO M., MATSUYAMA T., MORI H., 1979b, « Structural analysis of complex aerial photographs », *Proceeding IJCAI'79 Proceedings of the 6th international joint conference on Artificial intelligence,* vol. 2*,* pp.610-615.

NORBERT D., 1988 « From the balloon camera to the microprocessor-controlled LMK aerial survey camera system » *ISPRS Archives - Volume XXVII - Part B6 - Commission VI, XVIth Congress, Kyoto, Japan*, pp. 60-70.

NORDBECK S., 1971, « Urban allometric growth », *Geografiska Annaler. Series B, Human Geography*, 53, n° 1, pp. 54-67.

NUNINGER L., FRUCHART C., OPITZ R., 2010, « LiDAR : quel apport pour l'analyse des paysages ? », *AGER (Association d'étude du monde rural gallo-romain), Bulletin de liaison*, n° 20, décembre 2010, pp. 34 -43.

NUNINGER L., OSTIR K., 2009 « Rapport d'activité 2009°: 3ème année°: LEA franco-slovène ModeLTER », (Sous la direction de), *Rapport d'activité de l'European Laboratory for Modelling of Landscapes and Territories over the Long Term*, 17°p.

O

OPENSHAW S., 1984, *The modifiable areal unit problem*, Concepts and Techniques in Modern Geography 38, Geobooks, Norwich, UK: 41 p.

ORGANISATION MONDIALE DE LA SANTE (OMS), 1991, *Manuel d'épidémiologie pour la gestion de la santé au niveau du district*, Genève : 186 p. (Sous la direction de VAUGHAN J.P., MORROW, R.H.,)

_____, 2008, *Classification Internationale des Maladies pour l'Oncologie CIM-O-3*, Editions de l'OMS, Troisième Édition, Genève: 286 p. « Rédacteurs » FRITZ A., PERCY C., JACK A., SHANMUGARATNAM K., SOBIN L., PARKIN D., WHELAN S.,

P

PACIOREK CJ., LIU Y., 2009, « Limitations of Remotely Sensed Aerosol as a Spatial Proxy for Fine Particulate Matter », *Environmental Health Perspectives*, 117: 904-909.

POPESCU S., ZHAO K., NEUENSCHWANDER A., LIN C., 2011, « Satellite lidar vs. small footprint airborne lidar: Comparing the accuracy of aboveground biomass estimates and forest structure metrics at footprint level Original » *Remote Sensing of Environment*, vol. 115, pp. 2786-2797.

PROVENCHER L., DUBOIS J-M., 2007, *Précis de télédétection volume 4 : Méthodes de photo-interprétation d'image*, Presses de l'Université du Québec, Québec : 468 p.

PUISSANT A., 2003, *Information géographique et images à très haute résolution. Utilité et applications en milieu urbain*, Thèse de doctorat Université Louis Pasteur, Strasbourg I

PUISSANT A., LHOMME S., WEBER C., HE D.C., MORIN D., 2004, « Comparaison de la segmentation des images Ikonos et Quickbird pour l'identification du bâti en milieu urbain », *Revue Française de Photogrammétrie et de Télédétection*, n° 173/174, p. 157.

PUISSANT A., WEBER C., 2004, « Démarche orientée « objets-attributs » et classification d'images THRS », *Revue Française de Photogrammétrie et de Télédétection*, n° 173/174, pp. 123-134.

PUMAIN D., SAINT-JULIEN T., 2010a, *Analyse Spatiale : Les Interactions*, Armand Colin, Paris, 218 p.

_____., 2010b, *Analyse spatiale : les localisations*, Cursus géographie, Armand Colin, 2e édition, Paris : 190 p.

R

RALEIGH VS., BALARAJAN R., « Public health and the 1991 census », *BMJ,* vol. 309, pp. 287-288.

RASE W-D., 2001, « Volume-preserving interpolation of a smoth surface from polygon-related data », *Journal of Geographical Systems*, 3, pp. 199-213.

RAVAUD P., 2006, « L'épidémiologie clinique, base scientifique pour la prise en charge rationnelle des malades », *Rapport sur la science et la technologie de l'Académie des Sciences*, n° 23, pp. 249-259.

REVERCHON A., 2004, « Les grandes villes passées au crible: Une étude mesure et compare, pour la première fois, les performances des 200 plus importantes municipalités françaises en matière d'environnement et de cohésion sociale », *Le Monde - Développement Durable*, mercredi 10 novembre.

RICHARDS J., JIA X., 2006, *Remote Sensing Digital Image Analysis An Introduction*, 4th Edition, Springer, Germany : 439 p.

ROBINSON A., 1955, « The 1837 maps of Henry Drury Harness » *The Geographical Journal*, 121, n° 4 pp. 440-450.

RHODAIN F. et PEREZ C., 1985, *Précis d'entomologie médicale et vétérinaire*, Ed. Maloine, Paris

S

SANDERS L., 1989, *L'analyse statistique des données en géographie*, GIP RECLUS, Montpellier : 267 p.

SCHOLZ S., GRUBER M., 2009, « Radiometric and Geometric Quality Aspects of the Large Format Aerial Camera ULTRACAM XP », in *Proceeding ISPRS Hannover Workshop 2009: High-Resolution Earth Imaging for Geospatial Information*, Volume XXXVIII-1-4-7/W5, 2-5 juin, Hanovre, Allemagne

SCROPE G.P., 1833, *Principles of Political Economy, Deduced from the Natural Laws of Social Welfare, and Applied to the Present State of Britain*, Longmans, London, U.K. : 457 p.

SERRA J., 1982*, Image analysis and mathematical morphology*, volume 1, Academic Press, London UK : 610 p.

SING T., SANDER O., BEERENWINKEL N., LENGAUER T., 2005, « ROCR: visualizing classifier performance in R », *Bioinformatics*, vol. 21, pp. 3940-3941.

SOCIETE AERODATA FRANCE, 2009, « Acquisition et prétraitement de données altimétriques par laser aéroporté sur les secteurs de Besançon et Mandeure-Mathay ». Rapport de synthèse à destination de la MSHE Claude Nicolas Ledoux-USR 3124 CNRS : 45 p.

SOHN G., DOWMAN I., 2007, « Data fusion of high-resolution satellite imagery and LiDAR data for automatic building extraction », *ISPRS Journal of Photogrammetry & Remote Sensing*, vol. 62, pp. 43-63.

SUTTON P., 1997, « Modeling population density with night-time satellite imagery and GIS », *Computers, Environment and Urban System*, 31, n° 3/4, pp. 227-244.

T

TOBLER W., 1969, « Satellite Confirmation of Settlement Size Coefficients», *Area*, vol. 1, n° 3, pp. 30-34.

_____, 1979, « Smooth pycnophylactic interpolation for geographical regions », *Journal of the American Statistical Association*, 74, n° 367, pp. 519-530.

TOURANCHEAU J., avant propos, in WORONOFF A-S., DANZON A., 2008, *Le Cancer en Franche-Comté : Incidence et mortalité de 1980 à 2005,* Rapport de l'Observatoire Régional de la Santé de Franche-Comté et du registre des tumeurs du Doubs : 96 p.

TRAN A., GARDON J., POLIDORI L., 2004, « Application of Remote Sensing for Disease surveillance in urban and suburban areas », In: *Evidence-based practice manual : Research and outcome measures in health and human services*, Oxford University Press, New York, pp. 368-378.

TRAN A., GARDON J., WEBER S., POLODORI L., 2002, « Mapping disease incidence in suburban areas using remotely sensed data », *American Journal of Epidemiology*, vol. 152, pp. 662-668.

TRAN A., GOUTARD F., CHAMAILLE L., BAGHDADI N., LO SEEN D., 2010, « Remote sensing and avian influenza: a review of image processing methods for extracting key variables affecting avian influenza virus survival in water from earth observation satellites », *International journal of applied earth observation and geoinformation*, vol. 12, pp. 1-8.

TRICART J., RIMBERT S., LUTZ G., 1970, *Introduction à l'utilisation des photographies aérienne en géographie, géologie, écologie, aménagement du territoire. Tome I Notions générales, données structurales, géomorphologie*, Société d'édition d'enseignement supérieur, Paris : 247 p.

TUCKER C. J., SELLERES P.J, 1986, « Satellite remote sensing of primary production », *International Journal of Remote Sensing*, vol. 7, n° 11, pp. 1395-1416.

U

UPEGUI E., 2009, *Vers la mise à jour de la base de données « GEOPOLIS » en utilisant la morphologie mathématique pour extraire les contours des agglomérations ayant entre 10 000 et 500 000 habitants à partir des images Google Earth*, Rapport de stage, Master 2 - TGAE, Université Paris Diderot - Paris VII : 53 p.

V

VACHIER C., 1995, *Extraction de caractéristiques, segmentation d'image et morphologie mathématique,* Thèse de doctorat, Ecole Nationale Supérieure des Mines de Paris, 227 p

VAN DER SANDE C.J., DE JONG S.M., DE ROO A.P.J., 2003, « A segmentation and classification approach of IKONOS-2 imagery for land cover mapping to assist flood risk

and flood damage assessment », *International Journal of Applied Earth Observation and Geoinformation*, n° 4, pp. 217-229.

VEREGIN H., TOBLER W., 1997, « Allometric relationships in the structure of street-level databases », *Computers, Environment and Urban System*, 21, n° 3/4, pp. 277-290.

VIEL J-F., CLÉMENT M-C., HÄGI M., GRANDJEAN S., CHALLIER B., DANZON A., 2008, « Dioxin emissions from a municipal solid waste incinerator and risk of invasive breast cancer: a population-based case-control study with GIS-derive exposure », *International Journal of Health Geographics,* 7:4.

VIEL J-F., LEFEBVRE A., MARIANNEAU P., JOLY D., GIRAUDOUX P., UPEGUI E., TORDO N., HOEN B., 2011, « Environmental risk factors for haemorrhagic fever with renal syndrome in a French new epidemic area », *Epidemiology and Infection, Cambridge University Press 2010*, 139, pp. 867-874.

VIEL J-F., TRAN A., 2009, « Remote-sensed population estimates at a fine spatial resolution in a European urban area », *Epidemiology*, 20, pp. 214-222.

VONBUN F., WEIGHTMAN J., WILSON P., ELSMORE B., 1977, « A Discussion on Methods and Applications of Ranging to Artificial Satellites and the Moon », *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 284, pp. 443-450.

W

WARD M., NUCKOLS J., WEIGEL S., MAXWELL S., CANTOR K., MILLER R., 2000, « Identifying Populations 0Potentially Exposed to Agricultural Pesticides Using Remote Sensing and a Geographic Information System », *Environmental Health Perspectives*, n° 1, pp. 5-12.

WEBER C., 1995, *Images satellitaires et milieu urbain*, Série géomatique, Hermes, Paris : 185 p.

WEBSTER C., 1996, « Population and dwelling unit estimates from space », *Third World Planning Review*, vol. 18, pp. 155-176.

WENG Q., 2010, *Remote sensing and GIS integration Theories, methods and application*, Mc Graw Hill : 397 p.

WORONOFF A-S., DANZON A., 2008, *Le Cancer en Franche-Comté : Incidence et mortalité de 1980 à 2005,* Rapport de l'Observatoire Régional de la Santé de Franche-Comté et du registre des tumeurs du Doubs : 96 p.

WRIGHT J. K., 1936, « A method of mapping densities of population: with Cape Cod as an example », *Geographical Review*, 26, n° 1, pp. 103-110.

WU S-S., QIU X., WANG L., 2005, « Population estimation methods in GIS and remote sensing : A review », *GIScience and Remote Sensing*, 42, pp. 58-74.

_____., 2006, « Using Semi-variance Image Texture Statistics to Model Population Densities », *Cartography and Geographic Information Science*, vol. 33, n°. 2, pp 127-140.

WU S-S., WANG L., QIU X., 2008, « Incorporating GIS Building Data and Census Housing Statistics for Sub-Block-Level Population Estimation », *The Professional Geographer*, vol. 60, n° 1, pp. 121-135.

X

XIE Y., 1995, « The overlaid network algorithms for areal interpolation problem », *Computers, Environment and Urban System*, 19, n° 4, pp. 287-306.

Z

ZAMBON F., DE SANCTIS M., AMMANNITO E., CAPRIA M., CAPACCIONI F., CARRARO F., FONTE S., FRIGERI A., MAGNI G., MARCHI S., PALOMBA E., TOSI F., BLEWETT D., RAYMOND C., RUSSELL C., TITUS T., 2012, « Classification of dawn vir hyperspectral data of vesta », *43rd Lunar and Planetary Science Conference*, 19-23 Mars, The Woodlands, Texas, Etats-Unis.

Annexes

Annexe 1 : Morphologie Mathématique

Extraits de,

UPEGUI E., 2009, *Vers la mise à jour de la base de données « GEOPOLIS » en utilisant la morphologie mathématique pour extraire les contours des agglomérations ayant entre 10 000 et 500 000 habitants à partir des images Google Earth*, Rapport de stage, Master 2 - TGAE, Université Paris Diderot - Paris VII : 53 p.

Transformations ensemblistes classiques

Si on a deux ensembles, X (a dans la figure annexe 1-1) et Y (b dans la figure annexe 1-1), les opérations classiques dont on dispose sont : l'union ($X \cup Y$; figure annexe 1-1c) ; l'intersection ($X \cap Y$; figure annexe 1-1d) ; le complémentaire ($(X^c)_Z = X^c \cap Z$; figure annexe 1-1e) ; et la différence symétrique (X/Y ; figure annexe 1-1f).

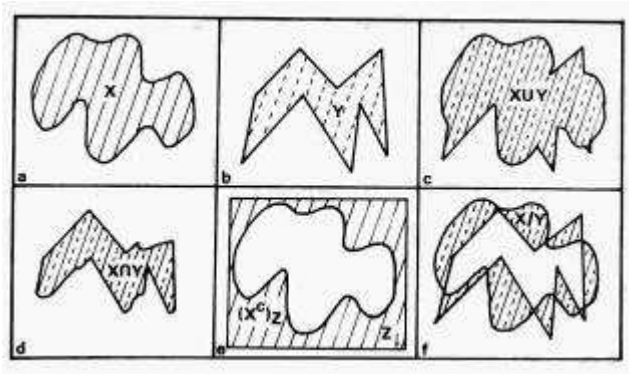


Figure annexe 1-1 Exemples de transformations ensemblistes classiques pour deux ensembles X et Y. Source : M. COSTER

Transformation en tout ou rien par un élément structurant

Pour effectuer une transformation en tout ou rien, « on doit choisir un élément, B, de géométrie connue, ensuite cet élément est déplacé de façon à ce que son origine passe par toutes les positions de l'espace (ensemble de points qui constituent un objet). Pour chaque position, on pose une question relative à l'union, à l'intersection ou à l'inclusion de B avec X ou dans X. La réponse sera positive ou négative. L'ensemble des points correspondant à des réponses positives forme un nouvel ensemble qui constitue l'image transformée »[58], laquelle sera stockée comme une image binaire.

[58] p. 67 in Précis d'analyse d'images. M. COSTER.

Chaque transformation en tout ou rien doit remplir quelques conditions euclidiennes ou géodésiques, et aussi qu'elles possèdent des propriétés algébriques et topologiques particulières[59]. Dans ce document on ne les approfondira pas. Nous utiliserons le disque comme élément structurant, compte tenu de ses propriétés d'isotropie et de convexité[60].

Transformation par érosion

Pour définir cette opération, l'élément structurant B qui parcourt toutes les positions dans l'espace R^2 doit répondre la question : est-ce que B_x est strictement inclus dans l'ensemble X ?, C'est à dire $B_x \subset X$? Dans le cas d'une réponse positive, une image transformée, représentée par un nouvel ensemble Y, sera constituée.

Ainsi, $Y = \{x : B_x \subset X \}$, on notera cette transformation $Y = E^B(X)$

L'érosion sert en particulier aux opérations de tri, en utilisant comme tamis B. Le résultat de cette transformation permet de séparer les surfaces des ensembles, et également de lisser les contours. D'autre part cette transformation permet aussi de créer un ensemble marqueur pour faire des reconstructions géodésiques.

Une autre approche pour comprendre cette transformation sur les images en teintes de gris (et non pas sur des images binaires) est d'« assimiler à un relief chaque pixel de la surface caractérisé non pas par une altitude, mais par une teinte de gris. On peut alors considérer que les valeurs maximales correspondent aux « pics » et les valeurs minimales aux des « vallées ».»[61] De ce fait, « l'érosion élargit les vallées et abaisse les pics de la fonction en teintes de gris »[62].

Transformation par dilatation

On peut définir cette opération de manière analogue à l'érosion ; dans chaque position de l'espace R^2, B doit répondre la question : B_x touche-t-il l'ensemble X ?, c'est à dire : $B_x \cap X \neq 0$? Ainsi, Y satisfait à l'équation $Y = \{ x : B_x \cap X \neq 0 \}$, on notera cette transformation comme $Y = D^B(X)$.

[59] Pour approfondir voir J.SERRA ou M.COSTER.
[60] p. 59 in ibid. J SERRA.
[61] p. 230 in SATFORM F. DEBAINE.
[62] p. 19 tableau A in Revue "PHOTO-INTERPRETATION". F. DEBAINE *et al*.

La dilatation permet de relier les parties des surfaces de l'ensemble écartées à moins de la taille de B, et également de lisser les contours. D'autre part, aussi « la dilatation comble les vallées et épaissit les pics de la fonction en teintes de gris[63].

Transformation par ouverture

Une fois que l'on connaît les transformations par érosion et par dilatation, on peut définir l'ouverture d'un ensemble X comme une érosion par un élément structurant B, suivi d'une dilatation sur l'ensemble érodé avec le même élément structurant B ; c'est à dire que $O^B(X) = D^B(E^B(X))$. « En général, après une ouverture, on ne retrouve pas l'ensemble de départ : l'ensemble ouvert $O^B(X)$, est plus régulier et moins riche en détails que l'ensemble X initial. La transformation par ouverture adoucit donc les contours, coupe les isthmes étroits, supprime les petits îles et les caps étroits »[64]. Autrement « l'ouverture rase les pics de la fonction en teinte de gris »[65].

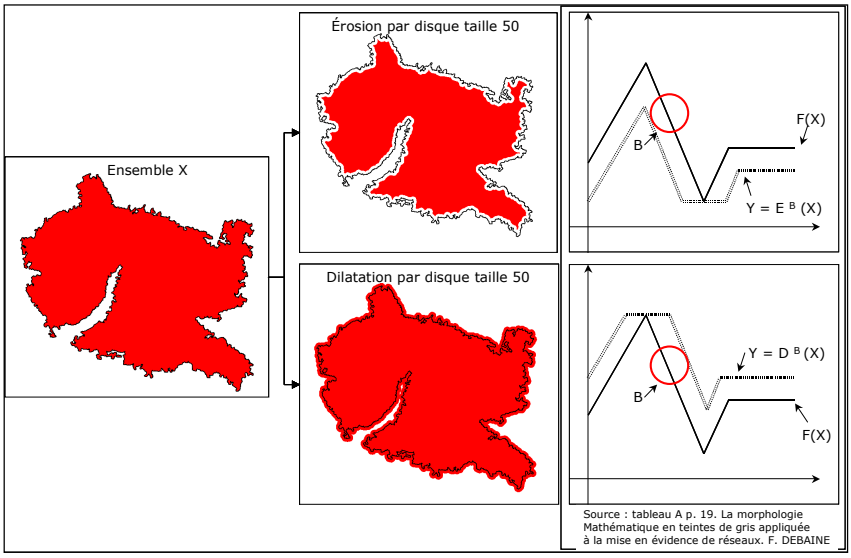


Figure annexe 1-2 Exemple des transformations érosion et dilatation par élément structurant disque de taille 50 sur l'ensemble X, dans ce cas : l'agglomération Porto-Novo au Bénin ; dans toutes les images le contour noir correspond à l'ensemble initial X.

Transformation par fermeture

De manière analogue, une fois définies l'érosion et la dilatation, on peut définir l'ouverture et la fermeture. Ainsi, la fermeture correspond à une dilatation d'un ensemble X

[63] Ibid F. DEBAINE *et al.*
[64] p. 83 in ibid. M. COSTER
[65] Ibid F. DEBAINE *et al.*

par un élément structurant B, suivi d'une érosion par le même élément structurant sur l'ensemble dilaté ; c'est à dire que $F^B(X) = E^B(D^B(X))$.

« Un ensemble fermé est également moins riche en détails que l'ensemble initial. La transformation par fermeture bouche les canaux étroits, supprime les petits lacs et les golfs étroits »[66]. Egalement « la fermeture comble les vallées de la fonction en teinte de gris »[67]

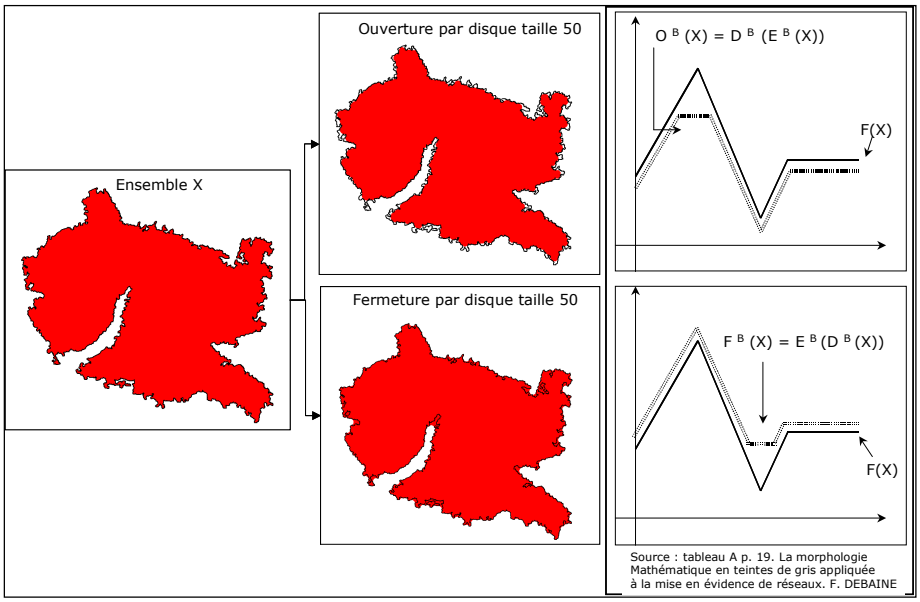


Figure annexe 1- 3Exemple des transformations ouverture et fermeture par élément structurante disque de taille 50 sur l'ensemble X, dans ce cas : l'agglomération Porto-Novo au Bénin ; dans toutes les images le contour noir correspond à l'ensemble initial X.

Transformation par reconstruction géodésique

Il s'agit d'une transformation morphologique qualifiée de géodésique car elle est basée sur les concepts géodésiques. Par géodésique, on entend le fait d'« être totalement inclus dans le ensemble X et correspondre au plus petit parcours entre deux points x1 et x2 »[68]. Pour y aboutir, ce type d'opération s'étend itérativement jusqu'à l'atteinte de la limite.

Ouverture par reconstruction géodésique[69]

Tout d'abord il faut différencier cette ouverture par φ(f) de l'ouverture algébrique employée jusqu'à maintenant. L'ouverture par reconstruction (figure annexe 1-4b) crée une

[66] p. 84 in ibid. M.COSTER.
[67] Ibid F. DEBAINE *et al.*
[68] p. 195 in ibid M. COSTER
[69] Nommé « ouverture morphologique » p. 371 in ibid. M. COSTER.

image transformée, à partir d'une image érodée, suivie d'une dilatation géodésique. Ainsi cette transformation peut être classée comme un filtre morphologique pour les images en teinte de gris.

Ce filtre $\varphi(f)$ garde les vallées et les petits lacs, et en général les valeurs sombres plus petites que la taille de l'ouverture, et conserve la forme des éléments après lissage.

Fermeture par reconstruction géodésique

De façon analogue, on notera cette fermeture par $\gamma(f)$, pour la raison exposée ci-dessus. De façon similaire, la fermeture par reconstruction géodésique (figure annexe 1-4c) crée une image transformée, à partir d'une image dilatée, suivie d'une érosion géodésique. Elle est également considérée comme un filtre morphologique.

Contrairement à l'ouverture $\varphi(f)$, cette transformation $\gamma(f)$, conserve les pics et les petites îles, et en général les valeurs claires inférieures à la taille de la fermeture, cependant elle lisse en conservant la forme des éléments.

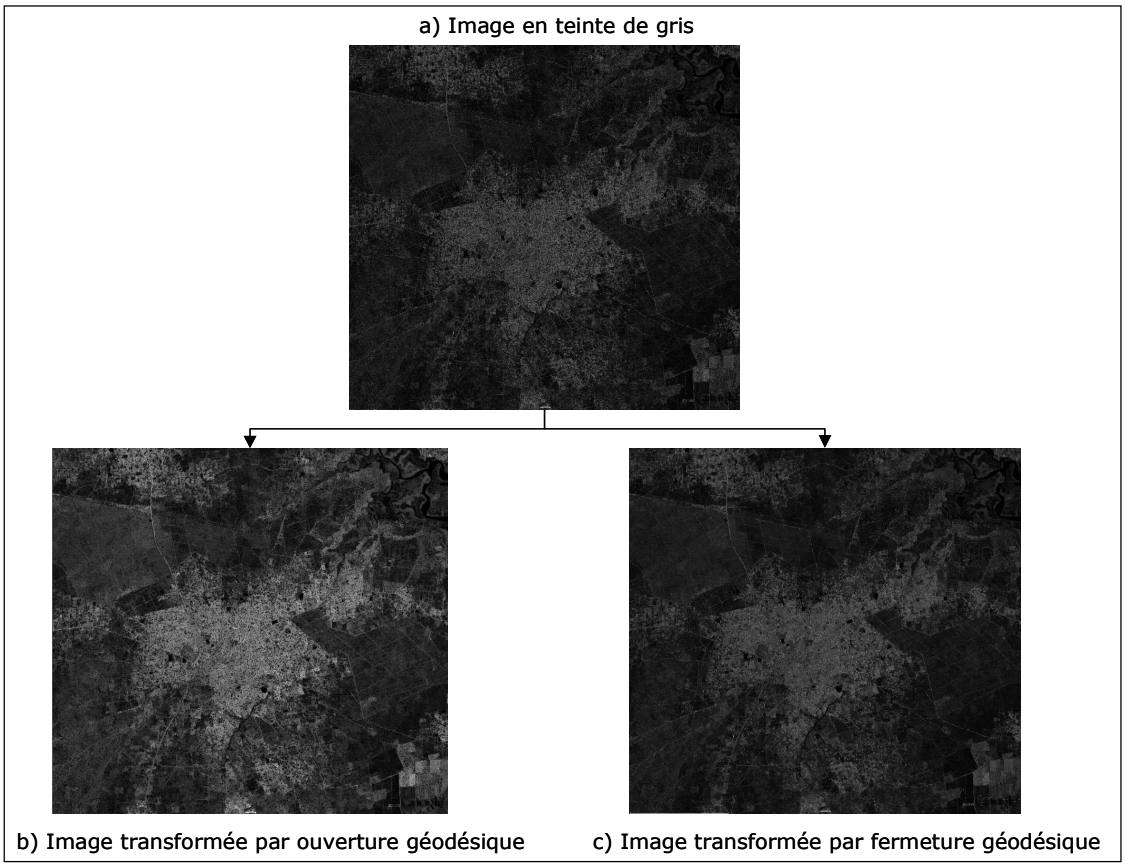


Figure annexe 1-4 Reconstructions géodésiques par un disque de taille 5, illustrées par une image transformée en teinte de gris à partir de l'agglomération de Brikama en Gambie.

BIBLIOGRAPHIE

COSTER M., CHERMANT J.L, (1989) *Précis d'analyse d'images*, Paris, Presses du CNRS.

DEBAINE F., (1990) « Extraction des réseaux linéaires a partir des images spot, exemple pris dans une région semi-aride : le nord-ouest de l'Inde », *ORSTOM - Journées Télédétection Bondy -Analyse quantitative des formes (SATFORM*) pp 225-239

DEBAINE F., MERING C., PONCET Y., (1988) « La morphologie mathématique en teintes de gris appliquée à la mise en évidence de réseaux », *Revue "Photo-interpretation" No 1988-5 fascicule 2*, pp 17 –26

DEBAINE F., FRANCFORT H. P., MERING C., (1989) « Analyse des images SPOT appliquée à la recherche archéologique au nord-ouest de l'inde : Recherche de linéaments » *Bulletin S.F.P.T n° 115,* pp 78-80

SERRA J., (1982) Image analysis and mathematical morphology – Vol I, London, Academic press.

VOIRON-CANICIO C. (1995) Analyse spatiale et analyse d'images par Morphologie Mathématique, Montpellier, Editions GIP RECLUS.

Annexe 2 : Résumé du nombre de classes identifiées lors des classifications, pour l'approche dirigée par régression logistique

Inclut la taille du pic granulométrie que chaque classe représente, ainsi que leur taille dans l'échantillon d'étude

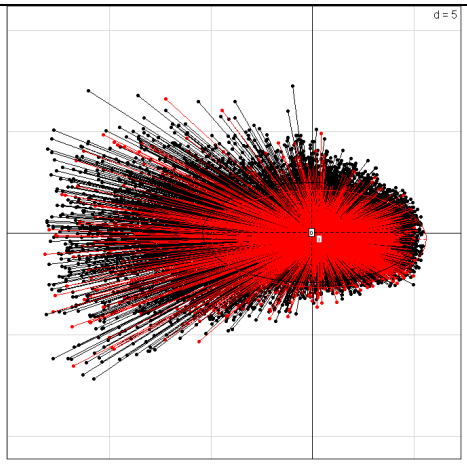
Niveau d'analyse	Nombre de l'itération de la classification	DPM $\gamma(f)$			DPM $\varphi(f)$		
		Nombre de la classe	**Taille granulo-métrique**	**Taille de l'échantillon d'étude**	**Nombre de classe**	**Taille granulo-métrique**	**Taille de l'échantillon d'étude**
Petit Bâtiment	itération 1	4	5	777	6	25	612
		8	15	440	7	15	1321
		NA			8	20	331
	itération 2	NA			6	20	615
	itération 3	4	10	604	5	15	708
		7	25	186	6	25	713
		NA			8	25	192
Bâtiment Moyen	itération 1	5	50	136	5	60	118
		6	30	139	6	50	185
		7	25	100	7	25	179
		8	60	112	8	40	183
		NA			9	50	100
	itération 2	1	25	50	4	30	154
		6	50	50	5	60	138
		7	60	50	6	50	184
		NA			7	40	254
		NA			8	25	195
		NA			9	30	145
	itération 3	5	40	154	4	40	397
		6	30	99	5	50	304
		7	60	95	6	25	251
		8	25	65	7	60	71
		9	40	79	8	40	202
		NA			9	40	55

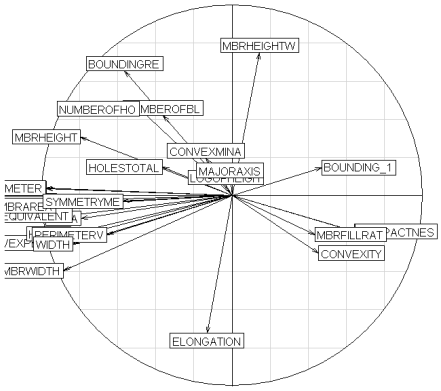
…Suite		DPM $\gamma(f)$			DPM $\varphi(f)$		
Niveau d'analyse	Nombre de l'itération de la classification	**Nombre de la classe**	**Taille granulo-métrique**	**Taille de l'échantillon d'étude**	**Nombre de classe**	**Taille granulo-métrique**	**Taille de l'échantillon d'étude**
Grand Bâtiment	itération 1	1	90	74	1	60	220
		5	70	124	4	80	99
		6	60	68	5	60	185
		7	100	89	7	60	50
		8	80	56	8	70	91
		9	70	50	9	80	50
	itération 2	1	70	50	2	80	50
		2	80	50	5	90	40
		8	60	58	9	100	40
		9	100	50	NA		
	itération 3	1	60	50	1	60	96
		5	70	66	4	70	74
		6	100	63	5	100	124
		7	60	50	6	80	115
		8	80	56	7	60	50
		9	70	55	8	80	50
		NA			9	70	50
Petit Bâtiment Orthophoto	itération 1	NA			6	25	50
		NA			7	15	17

Tableau annexe 2-1 Classes identifiées lors de la classification par régression logistique

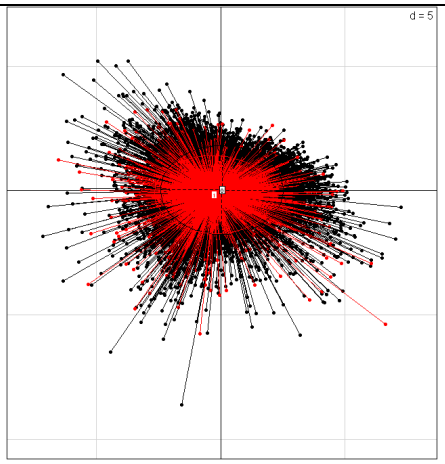
Annexe 3 : Représentation graphique des ACP et leur cercle de corrélation

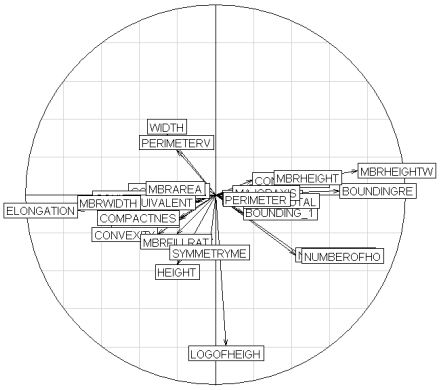
Figure annexe 3-1 Représentation graphique des ACP et leur cercle de corrélation du niveau d'analyse globale, 11 761 individus



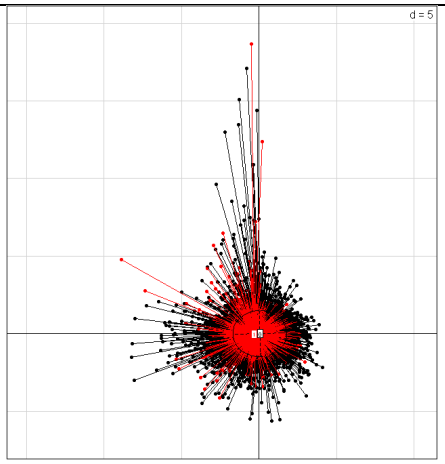


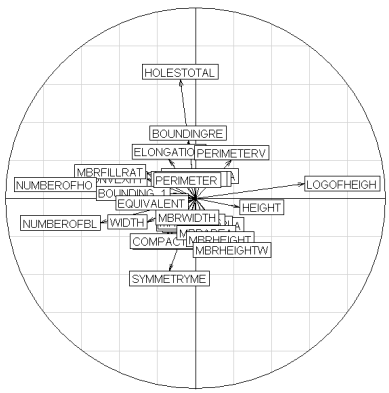
Plan principal 1-3 : Axe horizontal 1, axe vertical 3



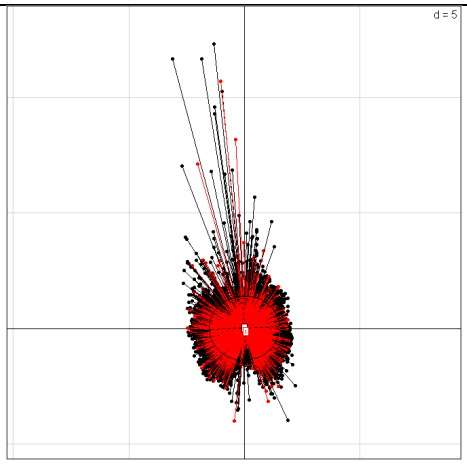


Plan principal 3-4 : Axe horizontal 3, axe vertical 4

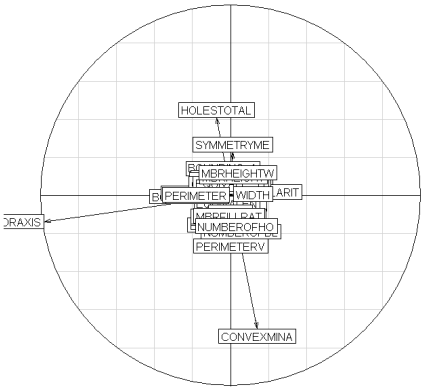


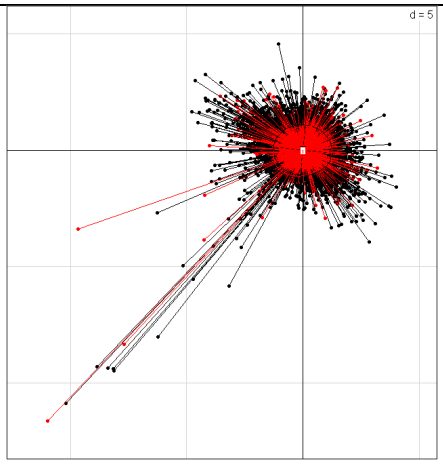


Plan principal 5-6 : Axe horizontal 5, axe vertical 6

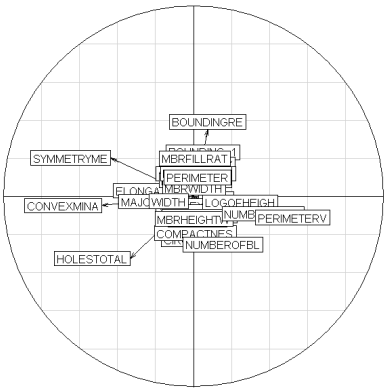


Plan principal 7-8 : Axe horizontal 7, axe vertical 8





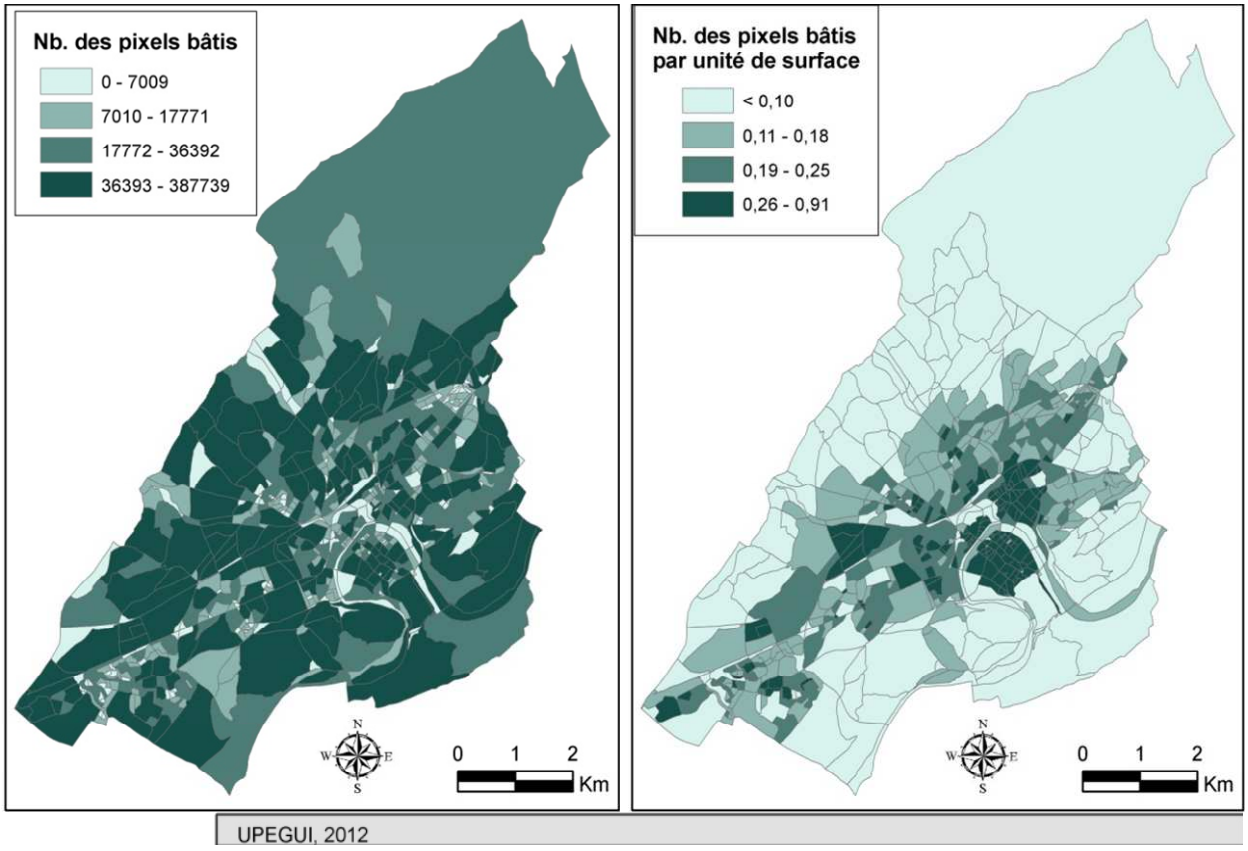
Plan principal 9-10 : Axe horizontal 9, axe vertical 10



Annexe 4 : Résultats de la modélisation de la population utilisant l'îlot comme unité géographique

Les mêmes analyses réalisées pour l'IRIS (chapitre 4), ont été effectuées sur les îlots (comme unité d'analyse pour l'échelle « d'arrivée ») pour valider le comportement de la méthode dasymétrique. La figure annexe 4-1 illustre les indicateurs de répartition pour la désagrégation de la population avec cette unité et en utilisant le bâti de l'IGN comme variable « bâti / non bâti » : à droite, la densité du bâti : et à gauche le pixel-bâti.

Figure annexe 4-1 : Indicateurs de répartition : à droite, la densité du bâti : et à gauche le pixel-bâti



En ce qui concerne l'indicateur du pixel-bâti, la distribution des unités géographiques ayant le plus grand nombre de pixels bâtis (dernier quartile) correspondent globalement à la dernière enveloppe de la ville. Néanmoins, l'unité géographique « Chailluz » se localise dans le troisième quartier au lieu du premier ; et ce en raison de la variation de taille, et donc d'échelle (plus détaillée), du reste des îlots.

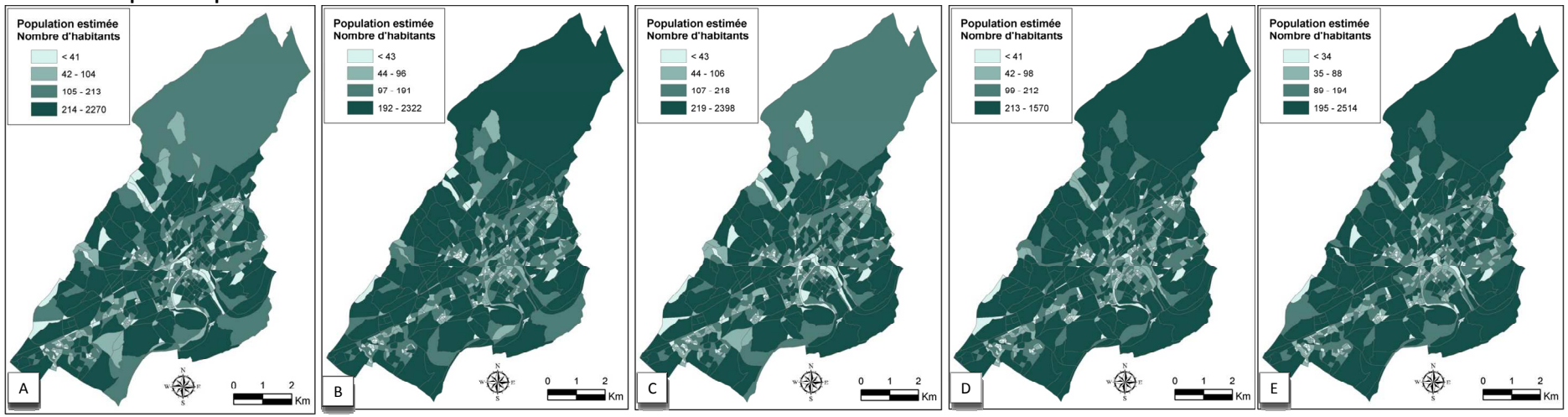
S'agissant de l'indicateur densité du bâti, il existe une concentration des densités du bâti vers l'intérieur de la ville, se prolongeant sur les axes routiers. Aussi, les îlots de la partie

extérieure de la ville (surtout les quartiers en habitat dispersé et dans la deuxième couronne sur la colline) atteignent-ils les valeurs de densités du bâti (premier quartile) les plus faibles.

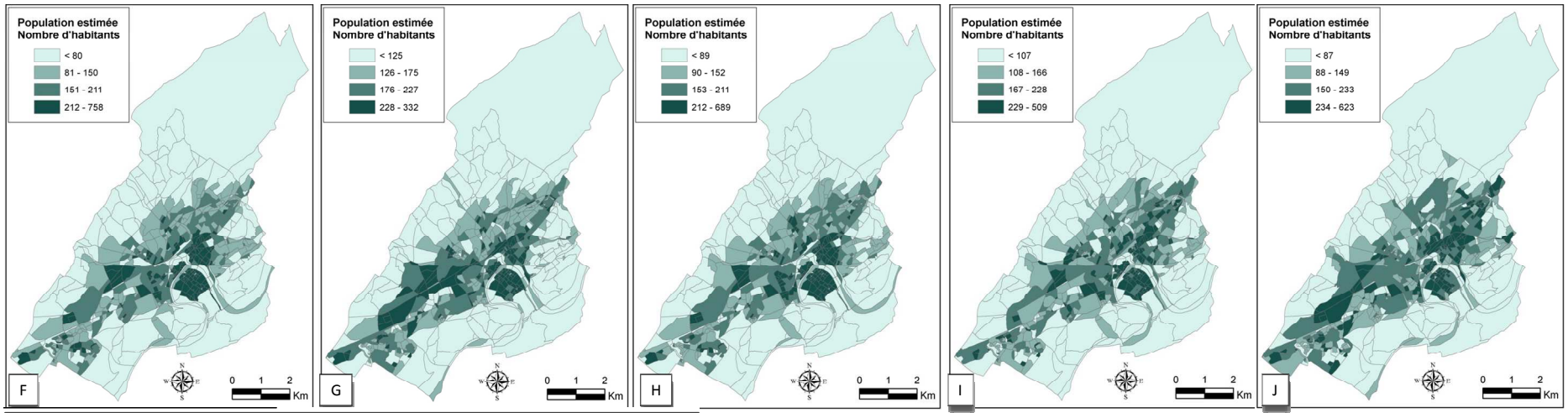
Les résultats de la modélisation par l'approche surface s'illustrent dans la figure annexe 4-2 (de A à E par MPBP, et de F à J par MDBP) et ceux de l'approche volume sont reportés dans la figure annexe 4-3 (de A à E par MPBP-3D, et de F à J par MDBP-3D). Les figures sont agencées, de gauche à droite, comme suit : bâti de l'IGN ; bâti extrait par ISODATA ; bâti extrait par classification hiérarchisée ; bâti extrait par classification ascendante hiérarchique, et bâti extrait par régression logistique. Comme lors de la modélisation des IRIS -tant pour l'approche surface que pour l'approche volume- les résultats obtenus pour les îlots se ressemblent à l'intérieur d'une même méthode de modélisation.

Figure annexe 4-2 : Population estimée par l'approche surface : de A à E par la méthode MPBP, de F à J par la méthode MDBP

Méthode du pixel bâti pondérée



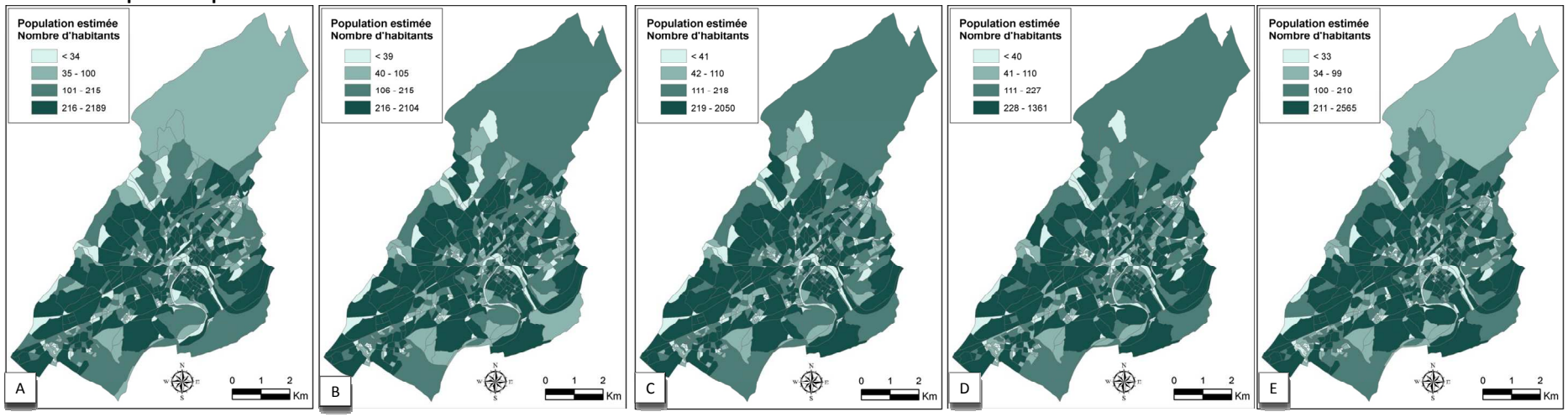
Méthode de la densité du bâti pondérée



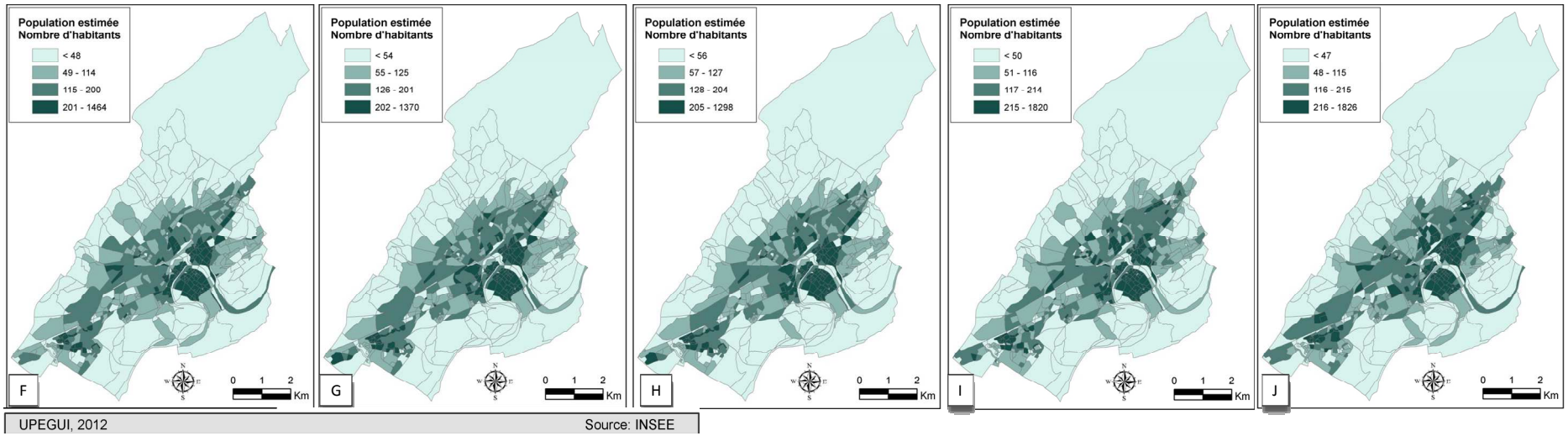
UPEGUI, 2012 Source: INSEE

Figure annexe 4-3 : Population estimée par l'approche volume. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D

Méthode du pixel bâti pondérée en 3D



Méthode de la densité du bâti pondérée en 3D



La validation des populations estimées par rapport aux populations du recensement est effectuée à l'aide de coefficients de corrélation intra-classe (CCI) dont les résultats sont reportés dans le tableau ci-dessous.

Tableau annexe 4-11 : Coefficient de corrélation intra-classe pour les îlots : N=680

Méthode de modélisation de la population		**Méthode de désagrégation**			
		Pixel-bâti CCI (IC95%)	**p**	**Densité du bâti** CCI (IC95%)	**p**
Approche par surface	Bâti IGN	0,30 (0,23-0,37)	4,37e-16	0,09 (0,01-0,16)	0,01
	C. ISODATA	0,18 (0,11-0,26)	5,87e-07	0,02 (-0,05-0,10)	0,26
	C. Hiérarchisée	0,30 (0,23-0,37)	5,77e-16	0,09 (0,00-0,16)	0,01
	C. Ascendante Hiérarchique	0,28 (0,21-0,35)	6,84e-14	0,06 (-0,01-0,13)	0,06
	C. Régression logistique	0,20 (0,13-0,27)	6,27e-08	0,09 (0,01-0,17)	0,008
Approche par volume	Bâti IGN	0,39 (0,32-0,45)	3,21e-26	0,17 (0,09-0,24)	5,32e-06
	C. ISODATA	0,43 (0,36-0,49)	6,57e-32	0,16 (0,09-0,24)	7,8e-06
	C. Hiérarchisée	0,44 (0,38-0,50)	1,22e-34	0,17 (0,10-0,24)	3,64e-06
	C. Ascendante Hiérarchique	0,47 (0,41-0,53)	4,15e-39	0,13 (0,05-0,20)	3,49e-04
	C. Régression logistique	0,39 (0,32-0,45)	5,62e-26	0,17 (0,10-0,24)	3,67e-06

En premier lieu, nous pouvons déduire que toutes les estimations fondées sur l'indicateur du « pixel-bâti » sont statistiquement significatives ($p<0,05$), de même que quasiment toutes les estimations fondées sur l'indicateur de la « densité du bâti » (sauf celles qui utilisent la classification ISODATA et la classification ascendante hiérarchique).

En deuxième lieu, les estimations qui utilisent l'indicateur du « pixel-bâti » obtiennent les valeurs les plus élevés du CCI (variant entre 0,18 et 0,47) ; tandis que pour celles utilisant l'indicateur de la « densité du bâti » les valeurs du CCI sont moindres (variant entre 0,02 et 0,17).

En troisième lieu, pour les deux indicateurs de répartition (pixel-bâti et densité du bâti), les valeurs atteintes par l'approche « volume » sont supérieures à celles obtenues par l'approche « surface ».
En dernier lieu, la classification hiérarchisée et la classification ascendante hiérarchique atteignent les valeurs les plus élevées tant dans les deux approches (« surface » et « volume ») que dans les deux méthodes de désagrégation (« pixel bâti pondéré » et « densité du bâti pondérée »), surpassant les valeurs du CCI obtenus par le bâti de l'IGN (utilisé comme donnée de référence). Toutefois, les valeurs du CCI obtenues par la classification ascendante hiérarchique dans la méthode de la densité du bâti pondérée ne sont pas statistiquement significatives ($p>0,05$).

L'analyse des diagrammes bi-variés portant sur les îlots permet d'effectuer deux constatations.

D'une part, la MDBP tend sous-estimer les populations (un grand nombre de points sont situés en dessous de la diagonale et sur la droite), notamment dans l'approche « surface » (figure annexe 4-4 de F à J) mais aussi dans l'approche « volume » (figure annexe 4-5 de F à J).

D'autre part, la MPBP présente une distribution de la population plus homogène même si certains points (entourés par des ellipses pointillées dans le graphique) sont soit surestimés (au dessus de la diagonale et sur la gauche), soit sous-estimés (en dessous de la diagonale et sur la droite). Les points surestimés correspondent aux zones d'utilisation commerciale, industrielle, ainsi qu'institutionnelle -comme par exemple Chamars ou le campus universitaire- tandis que les points sous-estimés appartiennent aux zones d'immeubles collectifs.

Figure annexe 4-4 : Diagrammes bi-variés des populations estimées par l'approche surface vs la population des îlot1999. De A à E par la méthode MPBP ; de F à J par la méthode MDBP

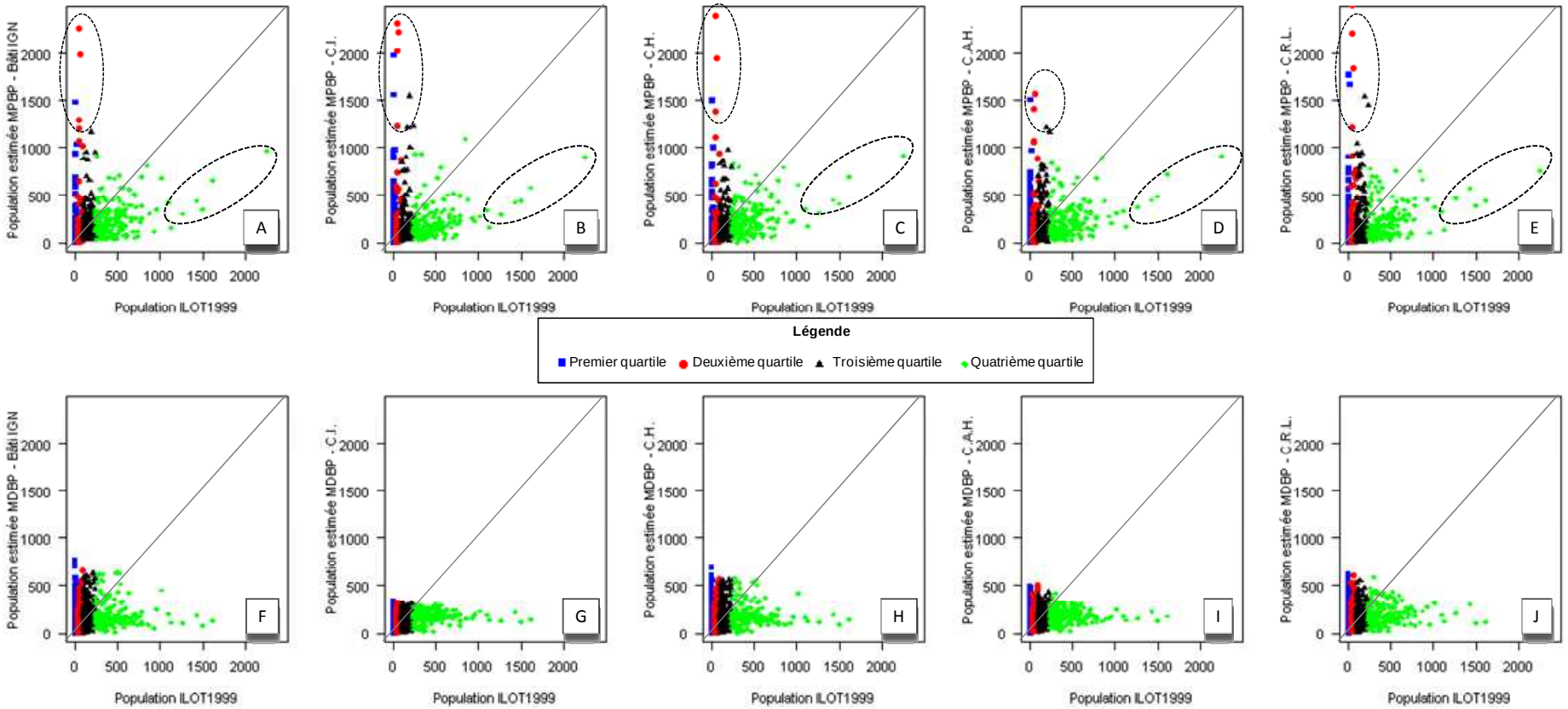
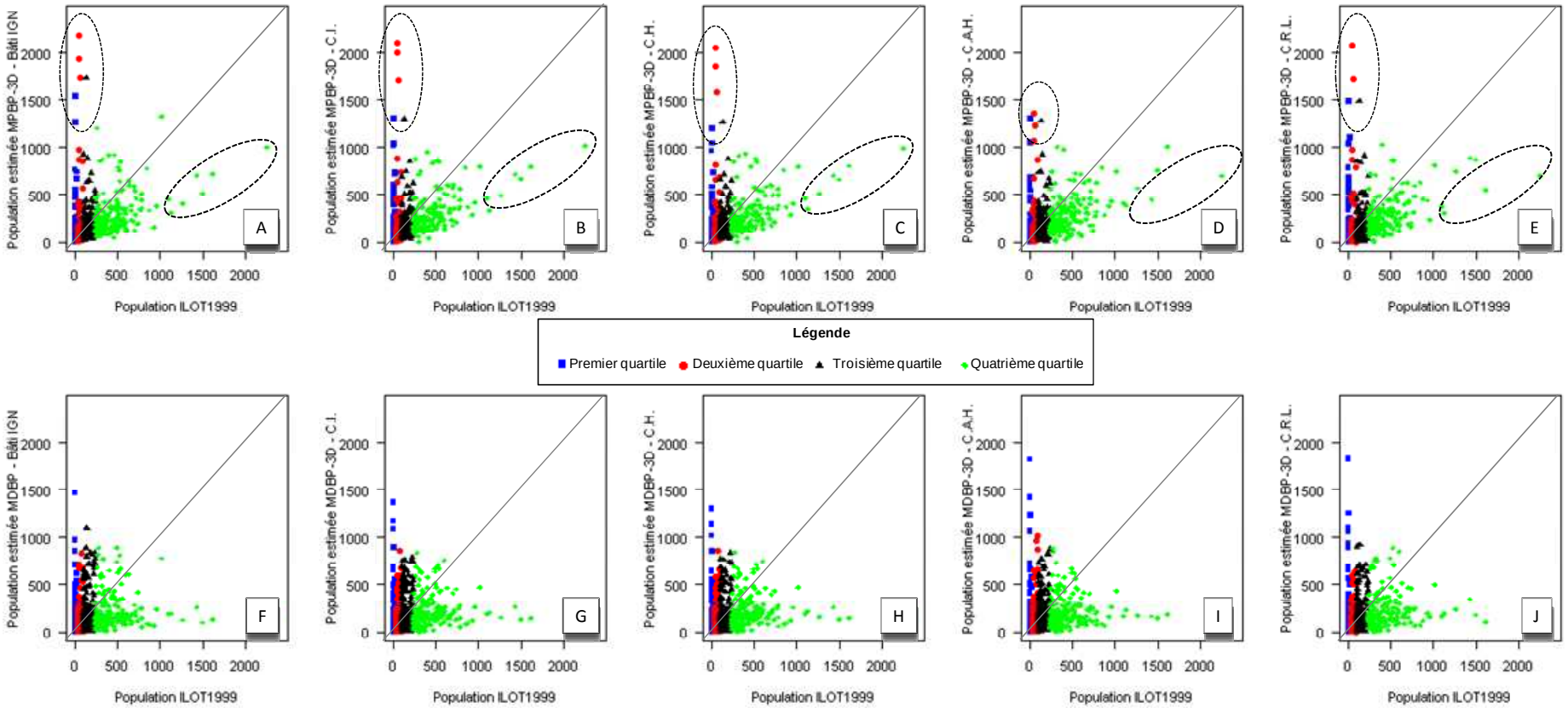


Figure annexe 4-5 : Diagrammes bi-variés des populations estimées par l'approche volume vs la population des îlot1999. De A à E par la méthode MPBP-3D ; de F à J par la méthode MDBP-3D



Dans l'analyse de sensibilité, 13 îlots ont été exclus en raison de l'utilisation non résidentielle des sols (4 aux Tilleroyes ; 1 à Châteaufarine ; 1 à Lafayette ; 1 à Chamars ; 1 à Saint Ferjeux ; 1 au Rosemont ; 2 à Bregille ; et 2 à l'Observatoire). Les résultats obtenus dans cette modélisation apparaissent dans le tableau annexe 4-2.

Tableau annexe 4-12 : Coefficient de corrélation intra-classe pour les îlots : N=667

Méthode de modélisation de la population		**Méthode de désagrégation**			
		Pixel-bâti CCI (IC95%)	**p**	**Densité du bâti** CCI (IC95%)	**P**
Approche par surface	Bâti IGN	0,47 (0,41-0,53)	3,42e-38	0,09 (0,01-0,16)	0,01
	C. ISODATA	0,30 (0,23-0,36)	2,69e-15	0,02 (-0,05-0,10)	0,24
	C. Hiérarchisée	0,47 (0,41-0,53)	1,86e-38	0,08 (0,00-0,16)	0,01
	C. Ascendante Hiérarchique	0,37 (0,31-0,44)	9,67e-24	0,06 (-0,01-0,13)	0,06
	C. Régression logistique	0,34 (0,27-0,40)	2,75e-19	0,10 (0,02-0,17)	0,005
Approche par volume	Bâti IGN	0,60 (0,55-0,64)	2,37e-66	0,17 (0,09-0,24)	5,46e-06
	C. ISODATA	0,62 (0,57-0,67)	2,01e-73	0,16 (0,09-0,24)	9,25e-06
	C. Hiérarchisée	**0,63** (0,58-0,67)	1,67e-75	0,17 (0,10-0,24)	4,43e-06
	C. Ascendante Hiérarchique	0,60 (0,55-0,65)	2,78e-67	0,13 (0,05-0,20)	4,53e-04
	C. Régression logistique	0,60 (0,55-0,65)	6,22e-68	0,17 (0,10-0,24)	3,72e-06

Ces résultats confirment, les conclusions découlant de l'analyse de résultats des IRIS. Ainsi, les valeurs du CCI restent quasiment identiques (après suppression des différents îlots dans le modèle) pour MDBP (à savoir, une entre 0,02 et 0,17), que ce soit par l'approche « surface » ou par l'approche « volume ». De même, ces résultats sont plus faibles que ceux obtenus par MPBP.

De façon analogue, on remarque une augmentation significative des valeurs du CCI obtenues par MPBP (passant d'un intervalle de [0,18 - 0,30] à un intervalle de [0,30 - 0,47]) ainsi que

de celles obtenues par MPBP-3D (augmentant d'un intervalle de [0,39 - 0,44] à un intervalle de [0,60 - 0,63]).

De même, hormis l'estimation effectuée en utilisant le « bâti » extrait par classification ISODATA avec l'approche « surface », en particulier par MDBP, toutes les estimations sont statistiquement significatives ($p < 0,05$).

En résumé, les performances des modélisations en fonction des bâtiments employés, sont par ordre decroissant d'importance, : la classification hiérarchisée (CCI = 0,47 par MPBP; CCI = 0,63 par MPBP-3D) ; le bâti de l'IGN (CCI = 0,47 par MPBP; CCI = 0,60 par MPBP-3D) ; la classification ascendante hiérarchique (CCI = 0,37 par MPBP; CCI = 0,60 par MPBP-3D) ; la classification par régression logistique (CCI = 0,34 par MPBP; CCI = 0,60 par MPBP-3D) ; et la classification par ISODATA (CCI = 0,30 par MPBP; CCI = 0,60 par MPBP-3D).

Une seconde analyse de sensibilité a été réalisée avec la méthode de désagrégation du pixel-bâti pondérée, excluant les îlots écartés dans l'analyse précédente. De manière analogue à l'analyse réalisée sur les IRIS notre zone d'étude a été divisée en trois groupes, à savoir : densité basse, densité moyenne et densité haute (tab. annexe 4-3). Puis, la population a été recalculée à l'intérieur de chaque groupe, ainsi que la valeur du coefficient de corrélation intra-classe. Ces résultats apparaissent dans le tableau annexe4-4.

Tableau annexe4-13 : Distribution des unités géographiques en fonction de la densité de la population

Unité géographique	**Densité basse** Habitantes /Nb. unité	**Densité moyenne** Habitantes/Nb. unité	**Densité haute** Habitantes /Nb. unité
Ilot	11 211 hab./222 îlots	33 114 hab./222 îlots	72 486 hab./223 îlots

Les valeurs moins élevées du CCI retrouvées pour la densité moyenne, par rapport aux autres densités, s'expliquent par la présence d'îlots ayant une population égale à zéro malgré la présence de bâtiments (voir 2.3.1, note de bas de page 22). Au-delà, les valeurs du CCI obtenus pour les trois densités différentes présentent un degré d'association fort entre la population observée et la population estimée par la méthode dasymétrique.

Tableau annexe4-14 : Coefficient de corrélation intra-classe pour les îlots en fonction de la densité de la population

Méthode de modélisation de la population		**Méthode de désagrégation du pixel-bâti pondérée**		
		Densité basse CCI (IC95%) N= 222	**Densité moyenne** CCI (IC95%) N= 222	**Densité haute** CCI (IC95%) N= 223
Approche par surface	Bâti IGN	0,51 (0,41-0,60)	0,85 (0,80-0,88)	0,75 (0,68-0,80)
	C. ISODATA	0,39 (0,27-0,49)	0,86 (0,82-0,89)	0,9 (0,87-0,92)
	C. Hiérarchisée	0,54 (0,44-0,63)	0,85 (0,80-0,89)	0,78 (0,73-0,83)
	C. Ascendante Hiérarchique	0,59 (0,5-0,67)	0,87 (0,83-0,89)	0,86 (0,82-0,89)
	C. Régression logistique	0,52 (0,41-0,61)	0,81 (0,76-0,85)	0,83 (0,78-0,86)
Approche par volume	Bâti IGN	0,46 (0,35-0,56)	0,77 (0,71-0,82)	0,69 (0,62-0,75)
	C. ISODATA	0,44 (0,33-0,54)	0,81 (0,75-0,84)	0,81 (0,76-0,85)
	C. Hiérarchisée	0,49 (0,38-0,58)	0,82 (0,77-0,86)	0,82 (0,77-0,86)
	C. Ascendante Hiérarchique	0,55 (0,46-0,64)	0,80 (0,75-0,84)	0,76 (0,70-0,81)
	C. Régression logistique	0,46 (0,35-0,56)	0,80 (0,75-0,84)	0,77 (0,71-0,82)

Annexe 5 : Prix A'Doc 2012

L'article intitulé « Utilisation des données télédétection à très haute résolution spatiale pour l'estimation de la population dans le cadre d'études épidémiologiques », a reçu le prix A'Doc 2012 (2ème lauréat) de la jeune recherche en Franche-Comté pour l'Ecole doctorale Langages, Espaces, Temps, Sociétés (LETS - ED No. 38). Cet article repose sur l'esprit interdisciplinaire de notre recherche. Il présente les différentes méthodes de classification (ISODATA, la classification hiérarchisée et la classification par régression logistique) employées pour extraire les bâtiments, puis pour modéliser la population (tant avec l'approche surface qu'avec l'approche volume). Les populations estimées ont été utilisées à leur tour pour calculer des taux bruts d'incidence du cancer de sein. Dans le cadre de la valorisation des travaux de recherche des doctorants, l'article est publié aux Presses universitaires de Franche-Comté (pp. 113-130 - ISBN 978-2-84867-427-8).

À travers ce recueil, le lecteur a la possibilité de développer son savoir dans de nombreuses disciplines, de penser des transferts, de dresser des parallèles entre des méthodes, des approches et des pratiques *a priori* très éloignées. Cette pluridisciplinarité est le reflet de la jeune recherche en Franche-Comté et de l'association A'Doc.

Lucie Bettinger, Géographie et aménagement des territoires,
Laboratoire ThéMA (UMR 6049)
La fermeture des paysages : un phénomène aux dimensions multiples, un défi pour les parcs naturels régionaux francs-comtois

Arnaud Leclerc, Chimie,
Institut UTINAM (UMR 6213)
Méthode de la trajectoire adiabatique contrainte (CATM) pour résoudre l'équation de Schrödinger en physique quantique

Émilie Liboz, Mathématiques,
Laboratoire de mathématiques (UMR 6623)
Mettre de l'ordre dans les représentations d'un groupe, oui mais lequel ?

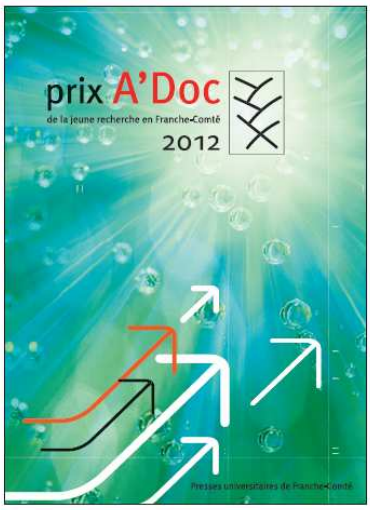
Leslie Mauchamp, Sciences de la vie et de l'environnement,
Laboratoire Chrono-Environnement (UMR 6249)
Structure spatiale et diversité des communautés végétales prairiales

Rayisa P. Moiseyenko, Sciences pour l'ingénieur,
Institut FEMTO-ST (UMR 6174)
Diffraction dans les cristaux phononiques

Javier Solano Martínez, Génie électrique,
Institut FEMTO-ST (UMR 6174)
Gestion d'énergie d'un véhicule hybride : approche par l'intelligence artificielle

Erika Upegui Cardona,
Géographie et aménagement des territoires,
MSHE Ledoux (USR 3124) - Laboratoire Chrono-Environnement (UMR 6249)
Utilisation des données télédétection à très haute résolution spatiale pour l'estimation de la population dans le cadre d'études épidémiologiques

Cette publication de l'association A'Doc est réalisée avec les soutiens de la Région Franche-Comté, de l'Université de Franche-Comté, des Presses universitaires de Franche-Comté et de la Ville de Besançon dans le cadre de la valorisation des travaux de recherche des doctorants.



Annexe 6 : Publication dans la revue « Photogrammetric Engineering & Remote Sensing »

L'article intitulé « GeoEye Imagery and Lidar Technology for Small-area Population Estimation: An Epidemiological Viewpoint » est paru en juillet (vol. 78, n°7, pp 693-702) dans la revue internationale *Photogrammetric Engineering & Remote Sensing,* publiée par la société américaine de photogrammétrie et télédétection (American Society for Photogrammetry and Remote Sensing).

GeoEye Imagery and Lidar Technology for Small-area Population Estimation: An Epidemiological Viewpoint

Erika Upegui and Jean-François Viel

Abstract

There is a continuing need for alternative approaches to obtain small-area populations when population census information is unavailable or inaccurate. The aim of this study was to develop an automated and exportable population model, based on GeoEye imagery and lidar technology, for estimating population counts and calculating disease rates in small areas. A satellite image (covering the City of Besançon, France) was processed to extract built-up pixels using ISODATA and hierarchical classifications. A dasymetric method was used to calculate population estimates at the block group level. Female breast cancer incidence rates were computed for each block group with the number of cases as numerator, and alternatively the population estimates and census data as denominators. A strong agreement was found between the estimated and the observed (census-based) incidence rates. This apportioning procedure could be of special interest to public health decision makers facing a lack of population census data.

Introduction

Adequate knowledge of the size and spatial distribution of human population is essential for determining disease rates and deriving meaningful indicators of health status, health services, and health systems for which population data are used as denominators. Traditionally, census data have been the primary source of information on population distribution and demographic characteristics to assess human exposures and risks to health outcome. However, some countries (and not only the "developing" countries) have difficulties conducting censuses, whereas others have no censuses at all. These difficulties are due to unaffordable costs, gaps in the civil registration system, extensive population movements (e.g., rural exodus), inaccessible remote areas, high illiteracy rates, high rates of population growth, and political troubles. In many cases, also, census tracts do not conform to, or nest within, the other spatial structures for which information is available, so that population data may need to be translated between different spatial structures for the purpose of data linkage and analysis (Briggs *et al.*, 2007).

Erika Upegui is with the House of Human Sciences and the Environment Claude Nicolas Ledoux, Besançon, France, and CNRS n° 6249 "Chrono-Environment," Faculty of Medicine, Besançon, France.

Jean-François Viel is with the Department of Public Health, Faculty of Medicine, 2 Avenue du Professeur Léon Bernard, 35043 Rennes, France, and formerly with CNRS n° 6249 "Chrono-Environment," Faculty of Medicine, Besançon, France (jean-francois.viel@univ-rennes.fr).

There is, therefore, a continuing need for alternative approaches to obtain small-area population denominators when population census information is unavailable, unreliable, or not available at the appropriate spatial resolution. Previous studies have examined the feasibility of using remote sensing to provide accurate estimates of population counts or assist with population interpolation (see Harvey, 2002a; Li and Weng, 2005, for summaries), assuming that correlation exists between remote sensing surrogates and population density.

Few attempts have been made to use these techniques for epidemiology, although studies (mainly focusing on vector-borne diseases) have increasingly used remotely sensed data for monitoring, surveillance, or risk mapping (Beck *et al.*, 2000). Tran *et al.* (2002) developed a method using remotely sensed data to estimate population density at a fine spatial resolution, and mapped the incidence of Q-fever in Cayenne and its suburbs (French Guiana). However, this is a specific tropical zone with numerous and heterogeneous landscape elements (composed of patches of urban areas mixed with natural areas such as mangrove forest, swamp, and tropical rain forest), and relative incidence rates were estimated instead of population counts. More recently, Viel and Tran (2009) described a Landsat ETM+-based population model for estimating fine-scale population data and characterizing high-incidence areas in an urbanized area (Besançon, France). However, sprawling induced some misclassification, and a saturation effect (remote-sensing estimates varied over a narrower range than the actual population density) induced some uncertainty in population count estimates (Viel and Tran, 2009). To address the saturation issue, a higher spatial resolution (rather than a higher spectral resolution) is required (Jensen and Cowen, 1999), but this was rarely available with previous satellite images.

The advent of very high spatial resolution satellite images and digital image processing techniques renewed the interest in using remote sensing to estimate human population counts. The 0.5-meter resolution of the GeoEye sensor seems sufficient to discriminate individual features (e.g., buildings, streets, and trees) within the urban mosaic. Two-dimensional building footprints provide a direct and accurate representation of the location of the people. However, they may not be sufficient to accurately estimate population counts, because spectral data alone capture little

Photogrammetric Engineering & Remote Sensing
Vol. 78, No. 7, July 2012, pp. 693–702.

0099-1112/12/7807–693/$3.00/0

Source :
http://digital.ipcprintservices.com/publication/?i=116511&p=&l=&m=&ver=&pp=
(consulté le 3 juillet 2012).

Liste des figures

Liste des tableaux

Table des matières

SENSORES REMOTOS Y EPIDEMIOLOGIA EN ZONA URBANA

Desde la extracción de edificaciones a partir de imágenes de muy alta resolución hasta la estimación de tasas de incidencia

RESUMEN

En epidemiología, un conocimiento preciso de las poblaciones en riesgo constituye un prerrequisito al cálculo de los indicadores de salud de una comunidad (tasa de incidencia). Sin embargo, los datos de la población pueden estar indisponibles, o ser poco fiables, o insuficientemente detallados por el uso epidemiológico.

El objetivo principal de este trabajo es obtener las tasas de incidencia en caso de ausencia de los datos demográficos, a una escala espacial infra comunal. Los objetivos secundarios son estimar las poblaciones humanas por intermedio de datos provenientes de sensores remotos de muy alta resolución espacial (MARE), evaluar el aporte de estos datos de MARE respecto a los datos de alta resolución espacial (Landsat) en un mismo marco urbano (Besançon), e implementar una metodología simple y robusta que garantice su exportabilidad a otras zonas.

Nosotros proponemos una solución en tres etapas, basada en la correlación existente entre la densidad de la población y la morfología urbana. La primera etapa consiste en extraer las edificaciones a partir de los datos de MARE. Estas edificaciones son utilizadas en la segunda etapa para modelar la población. Dicha población sirve de denominador, en la última etapa, para calcular las tasas de incidencia (cánceres). Los datos de referencia son utilizados en cada etapa para evaluar la efectividad de nuestra metodología.

Los resultados obtenidos resaltan el potencial de los sensores remotos para medir el estado de salud de una comunidad (bajo la forma de tasa bruta de incidencia) en una escala geográfica detallada. Dichas tasas estimadas pueden ser elementos de decisión para adaptar mejor la oferta de salud a las necesidades de salud de una comunidad, incluso en ausencia de los datos demográficos.

PALABRAS CLAVES: sensores remotos, muy alta resolución espacial, edificaciones, clasificación de imágenes, distribución de la población, epidemiología, tasas de incidencia, interdisciplinariedad.

TELEDETECTION ET EPIDEMIOLOGIE EN ZONE URBAINE

De l'extraction de bâtiments à partir d'images satellite à très haute résolution à l'estimation de taux d'incidence

RESUME

En épidémiologie, une connaissance précise des populations à risque constitue un pré requis au calcul d'indicateurs de l'état de santé d'une communauté (taux d'incidence). Néanmoins, les effectifs de population peuvent être indisponibles, ou peu fiables, ou insuffisamment détaillés pour un usage épidémiologique.

L'objectif principal de ce travail est d'obtenir des taux d'incidence en l'absence de données démographiques, à une échelle spatiale infra-communale. Les objectifs secondaires sont d'estimer les populations humaines par l'intermédiaire de données satellitaires à très haute résolution spatiale (THRS), d'évaluer l'apport de ces données THRS par rapport aux données à haute résolution spatiale (Landsat) dans un même cadre urbain (Besançon), et de mettre au point une méthodologie simple et robuste, pour garantir son exportabilité à d'autres zones.

Nous proposons une approche en trois étapes, fondée sur la corrélation existant entre la densité de population et la morphologie urbaine. La première étape consiste à extraire des bâtiments à partir des données télédétection THRS. Ces bâtiments sont utilisés dans la deuxième étape pour modéliser la population. A leur tour, ces populations servent de dénominateur, lors de la dernière étape, pour calculer des taux d'incidence (cancers). Des données de référence sont utilisées à chaque étape pour évaluer les performances de notre méthodologie.

Les résultats obtenus soulignent le potentiel de la télédétection pour mesurer l'état de santé d'une communauté (sous la forme de taux bruts d'incidence) à une échelle géographique fine. Ces taux d'incidence estimés peuvent alors constituer des éléments de décision pour mieux adapter l'offre de soins aux besoin de santé, même en l'absence de données démographiques.

MOTS-CLES : télédétection, très haute résolution spatiale, bâtiments, classification d'images, distribution de population, épidémiologie, taux d'incidence, interdisciplinarité.

REMOTE SENSING AND EPIDEMIOLOGY IN URBAN ZONE

From extraction of buildings from very high resolution satellite images to the estimation of incidence rates

ABSTRACT

In epidemiology, a precise knowledge of populations at risk is a prerequisite for calculating state of health indicators of a community (incidence rates). The population data, however, may be unavailable, unreliable, or insufficiently detailed for epidemiological use.

The main objective of this research is to estimate incidence rates, in cases of absence of demographic data, at an infra-communal scale. The secondary objectives are to estimate the human population through satellite data at very high spatial resolution (VHSR), to assess the contribution of this data (VHSR) compared with high spatial resolution data (Landsat) in a same urban framework (Besançon), and to develop a simple and robust methodology to ensure its exportability to other areas.

We proposed a three-step approach based on the correlation between population density and urban morphology. The first step is to extract buildings from VHSR imagery data. These buildings are then used in the second step to model the population data. Finally, this population data is used as the denominator to calculate incidence rates (cancers). Reference data are used at each step to assess the performance of our methodology.

The results obtained highlight the potential of remote sensing to measure the state of health of a community (in the form of crude incidence rates) at a fine geographical scale. These estimated incidence rates can be utilized as elements of decision to adapt better customized healthcare with respect to the health needs of a given community, even in the absence of demographic data.

KEY WORDS: remote sensing, very high spatial resolution, buildings, imagery classification, population allocation, epidemiology, incidence rates, interdisciplinarity.