GAME
LIEU
FNAU
IRSTV
LAVUE
LIENSs
LISST
LRA

Identification automatique d'une typologie urbaine des îlots urbains en France

Objectif de document	Détail de la mise au point de l'identification typologique automatique
Type de document	Livrable
Référence au projet	Tâche 1.3 : Analyse automatique des îlots urbains de France

Version	Date	Auteur	Modifications
v0	17/12/2015	Alexandre AMOSSÉ	
v1			

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Table des matières

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

A. Contexte de la mission

A.1. Le projet MApUCE

Le projet de recherche MApUCE (Modélisation Appliquée et droit de l'Urbanisme : Climat urbain et Énergie) vise à intégrer, dans les politiques urbaines et les documents juridiques les plus pertinents, des données quantitatives de microclimat urbain, climat et énergie, dans une démarche applicable à toutes les villes de France. Pour cela deux objectifs sont en ligne de mire :
(1) à partir de bases de données nationales, modéliser le microclimat urbain, la consommation d'énergie liée aux bâtiments et le comportement énergétique des habitants et usagers ;
(2) intégrer les données quantitatives de microclimat urbain, climat et énergie obtenues par modélisation dans les procédures juridiques et les politiques urbaines.

C'est dans le cadre du premier objectif du projet MApUCE que s'inscrit la tâche 1.3 ; tâche dont la finalité est de fournir certaines données d'entrée pour les modèles de microclimat urbain, de consommation d'énergie et de comportement énergétique. Ces données singulières consistent en la typologie des bâtiments des îlots urbains de France. En effet, connaître le type de bâtiment par une classification typologique est indispensable aux modèles du projet, car selon si le bâtiment est un pavillon, un immeuble ou un bâtiment d'activité, la consommation d'énergie, le comportement énergétique des usagers et donc le microclimat urbain seront différents.

A.2. Amélioration du modèle GENIUS

Le modèle GENIUS (Tornay *et al.*, 2015), développé par le LRA et le GAME dans le cadre du projet MUSCADE (Modélisation Urbaine et Stratégie d'adaptation au Changement climatique pour Anticiper la Demande et la production Énergétique) permet de transformer des bases de données urbaines existantes en des cartes archétypales. Ces cartes archétypales, d'une résolution de 250m par 250m, contiennent des informations sur la typologie, la géométrie, la matérialité, les équipements et les usages de chaque bâtiment. Ces données produites par GENIUS peuvent ensuite servir de données d'entrée au modèle TEB (Town Energy Balance) développé par le GAME afin d'évaluer le microclimat urbain et les consommations énergétiques des bâtiments.

Bien que l'utilisation de cet outil soit très intéressante pour établir des *scenarii* prévisionnels du climat urbain, ce modèle GENIUS n'a cependant été développé et utilisé que sur les villes de Paris dans le cadre du projet MUSCADE et de Toulouse dans le cadre du projet ACCLIMAT (Adaptation au Changement CLIMatique de l'Agglomération Toulousaine). Or il se trouve que le projet MApUCE vise une toute autre échelle : le territoire national français ; et pourquoi pas à plus long terme d'autres territoires nationaux. Cette différence d'échelle nécessite de nombreuses adaptations, car travailler à plus grande échelle signifie augmenter la durée des temps de calculs et de traitement de données, ce qui implique la mise en œuvre de moyens plus conséquents et plus performants. C'est pourquoi il est nécessaire d'automatiser plusieurs traitements analytiques, y compris la classification typologique, pour que le projet MApUCE puisse à terme fournir des données cohérentes, obtenues sans trop de délais d'attente, avec des méthodes dont la répétabilité est irréprochable.

L'autre différence majeure entre ACCLIMAT et MApUCE est la détermination des unités spatiales. Dans MapUCE, il ne s'agit pas de mailles de 250m sur 250m, mais d'îlots urbains déterminés par tesselation de Voronoï. Ces îlots urbains, appelés USR pour Unités Spatiales de Référence, peuvent donc être de taille variable et peuvent contenir zéro, un ou plusieurs bâtiments et/ou blocs de bâtiments (ensembles de bâtiments contigus).

Néanmoins, plusieurs éléments du modèle GENIUS ont inspiré MApUCE. Ces éléments/méthodes ont donc été repris et sont en cours d'amélioration : l'extraction des données urbaines de départ et leur traitement (dont la tesselation de Voronoï) se font à partir de données nationales et constituent la tâche 1.1 du projet MApUCE ; la matérialité, les équipements et les usages des bâtiments sont déterminés dans la tâche 1.2 ; la géométrie des bâtiments est constituée d'un ensemble d'indicateurs morphologiques calculés dans la tâche 1.4 à trois échelles distinctes : bâtiment, bloc de bâtiments et USR ; enfin l'identification automatique de la typologie, réalisée à partir de la géométrie des bâtiments, fait l'objet de la tâche 1.3.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

A.3. Evolution de la typologie urbaine

La typologie est un élément représentatif d'une classe d'individus qui écarte les individus atypiques pour obtenir une représentation abstraite de la réalité. Une typologie urbaine s'établit à partir de critères pouvant relever des groupes sociaux, de l'usage, de la morphologie,… Une première typologie urbaine a été développée par Marion Bonhomme pour le modèle GENIUS dans le cadre du projet MUSCADE, puis elle a été revisitée par Nathalie Tornay dans le cadre de la tâche 1.2 « Analyse architecturale des bâtiments typiques de France » du projet MApUCE.

Typologie développée par Bonhomme M. (2013)

Bonhomme (2013) a développé une typologie constituée de 7 quartiers types au sein du modèle GENIUS comme présenté dans la Figure 1.

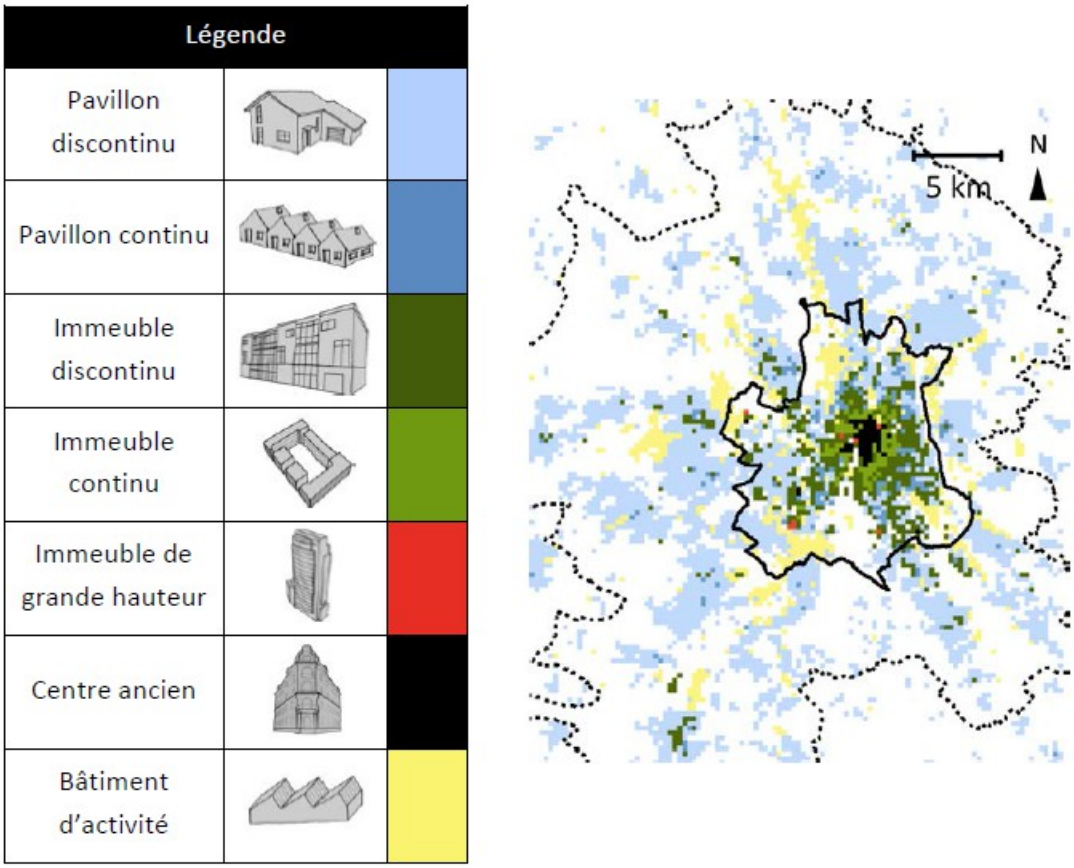


Figure 1 : Toulouse en 2010 selon GENIUS : répartition des types de quartier sur la ville (trait plein) et sur le pôle urbain (trait pointillé). (Bonhomme, 2013)

Pour cela des données provenant de la BD TOPO de l'IGN ainsi que de la BD IRIS de l'INSEE ont été transformées en indicateurs urbains à l'aide du SIG (Système d'Information Géographique) ArcGIS.
Deux premiers types ont alors pu être identifiés : le quartier type « bâtiment d'activités » correspondant aux bâtiments industriels, agricoles et commerciaux dont l'emprise au sol est supérieure à 0,6 fois l'emprise au sol des autres types de bâtiments ; le quartier type « immeuble de grande hauteur » correspondant aux quartiers où la hauteur moyenne des bâtiments est supérieure à 30m.
Après étude de corrélation, un certain nombre d'indicateurs urbains, présentés Tableau 1, ont été retenus. Une ACP (Analyse en Composantes Principales) a ensuite été réalisée sur ces variables via le logiciel TANAGRA.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Tableau 1 : indicateurs urbains utilisés pour l'ACP ayant permis la distinction des différents quartiers-types (Bonhomme, 2013).

Paramètres	
H_EcTy	Écart type des hauteurs des bâtiments dans la maille
Comp_EcTy	Écart type des compacités des bâtiments
Cont_EcTy	Écart type des contiguïtés des bâtiments
S_pl_moy_B	Surface de plancher moyenne par bâtiment
Orien_EcTy	Écart type des orientations des bâtiments
NbBat_Mail	Nombre de bâtiment par maille
H_moy_bat	Hauteur moyenne des bâtiments dans la maille
Comp_bat	Coefficient de compacité des bâtiments
Cont_bat	Coefficient de contigüité des bâtiments
S_env_ext	Surface d'enveloppe extérieure des bâtiments chauffés
Dens_br	Densité bâtie brute de la maille
CES	Coefficient d'emprise au sol
H_EcTy_bl	Écart type des hauteurs des bâtiments dans un même bloc bâti.
Spl_Moy_bl	Surface de plancher moyenne par bloc de bâtiment.
Comp_bloc	Coefficient de compacité des blocs
S_cours	Surface moyenne des cours
O_cours	Coefficient d'ouverture des cours
Nbbat_bl	Nombre de bâtiment par bloc
Dens_route	Densité surfacique de route
Angle_EcTy	Écart type des directions des routes
Larg_EcTy	Écart type des largeurs des routes
Long_route	Longueur route
Larg_route	Largeur moyenne des routes
Dens_veg	Densité surfacique de végétation
age_maj	Age majoritaire
Coll_Ind	Pourcentage de collectif et d'individuel
Dist_Centr	Distance au centre de la commune
H_Indus	Hauteur moyenne des bâtiments d'activités
Ssol_indus	Surface au sol des bâtiments d'activités dans la maille
D_hab_mail	Densité d'habitants approchée
D_hab_N	Densité d'habitants nette (par m² bâti)
D_hab_br	Densité d'habitants brute (par m² de maille)
Voit_men	Nombre de voiture personnelle par ménage

L'ACP est une méthode de réduction du nombre de variables permettant des représentations géométriques des individus (ici des îlots urbains) et des variables (ici des indicateurs urbains). Les proximités géométriques traduisent des associations statistiques entre individus (révélant des groupes d'individus qui se ressemblent) et entre variables (révélant des corrélations entre variables). L'ACP est une méthode dite factorielle car elle construit de nouvelles variables synthétiques, les composantes principales, en combinant les variables initiales. Les composantes principales décrivent les directions principales du nuage de point et sont le résultat d'une combinaison linéaire des variables initiales, ce qui fait de l'ACP une méthode d'analyse factorielle linéaire.

Cette ACP a ainsi permis d'identifier 5 quartiers types : « pavillonnaire discontinu », « pavillonnaire continu », « immeuble discontinu », « immeuble continu » et « centre ancien » (Figure 2). Les critères de classification de

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

la typologie au sein de GENIUS sont alors dépendant de la première composante principale de l'ACP, représentée par l'axe 1, comme exposé dans le tableau 2.

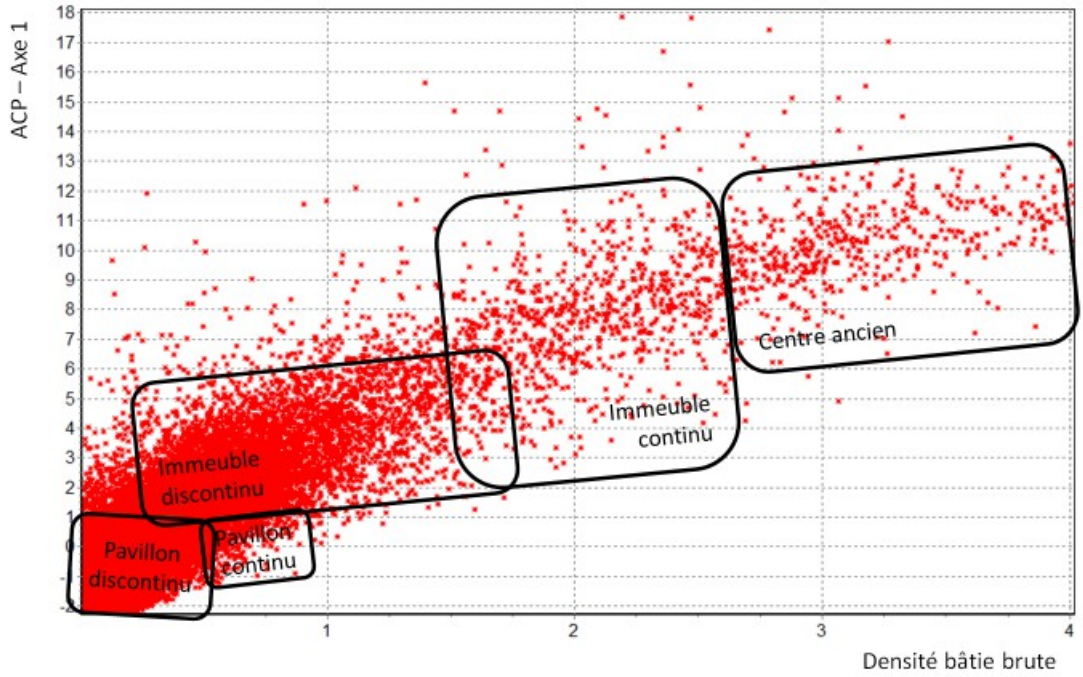


Figure 2 : Axe 1 de l'ACP en fonction de la densité bâtie brute (Bonhomme, 2013).

Tableau 2 : Critères de classification des quartiers types de GENIUS (Bonhomme, 2013)

Nom de l'ilot	Conditions
Pavillonnaire discontinu	Valeur sur l'axe 1 de l'ACP < 1 Densité brute < 0,6
Pavillonnaire continu	Valeur sur l'axe 1 de l'ACP < 1 Densité brute > 0,6
Immeuble discontinu	Valeur sur l'axe 1 de l'ACP > 1 Densité brute < 1,8
Immeuble continu	Valeur sur l'axe 1 de l'ACP > 1 1,8 < Densité brute < 2,5
Immeuble de grande hauteur	Hauteur moyenne > 30 m
Centre ancien	Valeur sur l'axe 1 de l'ACP > 1 Densité brute > 2,5
Bâtiment industriel ou commercial	Surface au sol indus / Surface au sol > 0,6
Non bâti	CES < 0,05

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Typologie revisitée par Tornay N. (2015)

Dans le cadre de la tâche 1.3 du projet MApUCE, la typologie développée par Bonhomme (2013) a d'abord été étoffée en ajoutant aux 7 quartiers types déjà définis 3 types intermédiaires : pavillon semi-continu, habitat intermédiaire et immeuble semi-continu. Pour valider cette typologie, il était nécessaire de vérifier la présence de ces 10 types de bâtiment au sein de plusieurs villes du territoire français. Une enquête a donc été menée auprès de professionnels de l'urbanisme par l'intermédiaire de la FNAU (Fédération Nationale des Agences de l'Urbanisme). Ainsi, en questionnant les professionnels de l'urbanisme sur les opérations existantes sur le territoire, l'enquête a permis de constituer un corpus d'études de cas à l'échelle de la France.

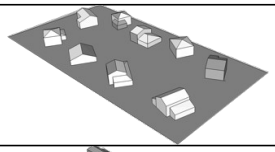
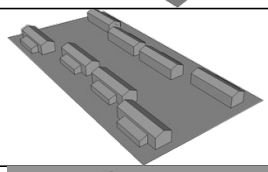
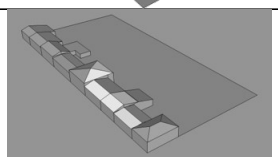
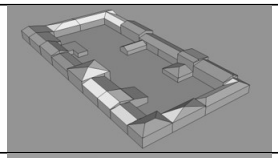
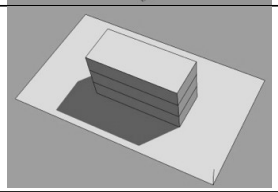
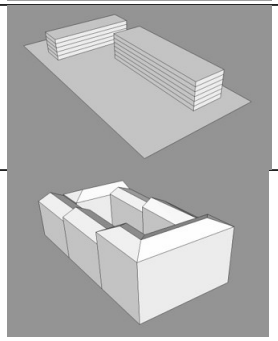
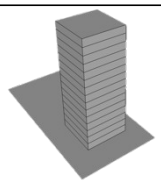
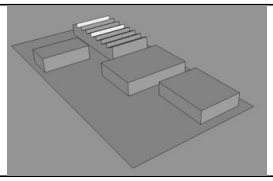
Après analyse des résultats d'enquête, Tornay (2015) a découvert que certains types, notamment « pavillon continu » et « immeuble continu », se déclinent selon 2 morphologies : îlot ouvert et ilot fermé ; et que d'autres types, comme « l'habitat intermédiaire », ne font pas l'objet d'un type particulier et se répartissent au sein des autres types. Tornay (2015) a donc finalement affiné la typologie urbaine comme exposé dans le tableau 3 et ainsi finalement obtenu la typologie décrite dans le tableau 4.

Tableau 3 : Synthèse des liens entre typologie étudiée et typologie finalement proposée (Tornay, 2015)

Typologies présentées lors de l'enquête	**proposition**
Pavillon discontinu	Pavillon discontinu
Pavillon semi-continu	Pavillon semi-continu
Pavillon continu	Pavillon continu sur îlot ouvert
	Pavillon continu sur îlot fermé
Habitat intermédiaire	Pavillon continu sur îlot ouvert
	Pavillon continu sur îlot fermé
	Immeuble continu sur îlot ouvert
	Immeuble continu sur îlot fermé
Immeuble discontinu	Immeuble discontinu
Immeuble semi-continu	Immeuble continu sur îlot ouvert
Immeuble continu	Immeuble continu sur îlot ouvert
	Immeuble continu sur îlot fermé
Immeuble de grande hauteur	Immeuble de grande hauteur
Centre ancien	Pavillon continu sur îlot fermé
	Immeuble continu sur îlot fermé
Bâtiment d'activité	Bâtiment d'activité

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Tableau 4 : Récapitulatif des caractéristiques des typologies du projet MApUCE (Tornay, 2015)

1	PAVILLON DISCONTINU		La typologie "pavillon discontinu" correspond aux îlots composés de bâtiments d'au moins quatre façades, RdC ou R+1, souvent implantés au centre de chaque parcelle.
2	PAVILLON SEMI CONTINU		La typologie "pavillon semi-continu" correspond aux bâtiments de type maison jumelée, implantés dans des lotissements ou cités jardin.
3	PAVILLON CONTINU SUR ILOT OUVERT		La typologie "pavillon continu ilot ouvert" correspond au bâtiment de type maison en bande, ou maison de ville, mitoyenne sur deux faces, en alignement sur rue.
4	PAVILLON CONTINU SUR ILOT FERME		La typologie "pavillon continu ilot fermé" correspond aux bâtiments à patio en bande, constructions en chartreuse dans les centres urbains, à l'habitat intermédiaire.
5	IMMEUBLE DISCONTINU		La typologie "immeuble discontinu" correspond aux bâtiments généralement implantés au centre de l'îlot avec quatre façades. Cela peut correspondre à des bâtiments tertiaires, du logement collectif ou des équipements recevant du public (ERP).
6	IMMEUBLE CONTINU SUR ILOT OUVERT		La typologie "immeuble continu sur îlot ouvert" correspond à un ensemble de bâtiments en partie alignés sur rue, cela peut correspondre à des bâtiments tertiaires, du logement collectif ou des ERP.
7	IMMEUBLE CONTINU SUR ILOT FERME		La typologie "immeuble continu sur îlot fermé" correspond aux bâtiments dont l'implantation constitue un îlot fermé. Cela peut correspondre aux centres historiques, aux tissus urbains de la révolution industrielle type immeubles haussmanniens.
8	BATIMENT DE GRANDE HAUTER		La typologie "immeuble de grande hauteur" correspond aux bâtiments d'au minimum 12 niveaux appelés aussi « tour », « barre d'immeuble » ou « gratte-ciel ».
9	BATIMENT D'ACTIVITE		La typologie " bâtiment d'activité " est représentée par les bâtiments industriels, commerciaux ou agricoles voire les équipements sportifs. Ils se caractérisent par leur emprise au sol et leur taille.
10	ILOT INFORMEL	–	La typologie « îlot informel » correspond aux constructions éphémères, non répertoriées par les cadastres.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Outre ces 10 types, un onzième type nommé « local annexe » a par la suite été rajouté. Bien que non dominant à l'échelle d'un quartier, les locaux annexes tels que garage, abri de jardin, local poubelles, véranda, kiosque, …, sont omniprésents et répertoriés par les cadastres. Il était donc indispensable de prendre en compte ces petits bâtiments, car aussi petits soient-ils, ils font partie intégrante des données et doivent être attribués à un type. La typologie urbaine du projet MapUCE est ainsi composée de 11 types de bâtiments :

- ba : bâtiment d'activité
- bgh : bâtiment de grande hauteur
- icif : immeuble continu sur îlot fermé
- icio : immeuble continu sur îlot ouvert
- id : immeuble discontinu
- info : construction informelle
- local : local annexe
- pcif : pavillon continu sur îlot fermé
- pcio : pavillon continu sur îlot ouvert
- pd : pavillon discontinu
- psc : pavillon semi-continu

C'est pour pouvoir identifier automatiquement cette typologie que nous allons développer un modèle prédictif à partir d'une méthode de classification supervisée.

B. Données et outils de classification

B.1. Méthodes de classification supervisée

Une méthode de classification permet de classer des objets/individus dans des classes à partir de variables qualitatives ou quantitatives caractérisant ces objets/individus. Il existe 2 types de méthodes de classification : les méthodes « non supervisées » et les méthodes « supervisées ».

Les méthodes non supervisées classent les individus sans connaissance *a priori* sur leur classement. Ces méthodes permettent ainsi de déterminer de manière objective quels individus se ressemblent et combien de classes/groupes distincts il est possible de réaliser. Cela permet également de poser des règles de classification si les groupes se distinguent assez bien selon les variables.

A l'inverse, les méthodes supervisées classent les individus selon des classes connues *a priori*. Ces méthodes se décomposent en 2 phases : une phase d'apprentissage puis une phase de prédiction. La phase d'apprentissage consiste à créer des règles de classification à partir d'individus dont on connaît à la fois les variables et le classement. Une fois ces règles établies, notre méthode de classification peut alors classer un nouvel individu à partir des seules variables qui lui sont propres. Deux situations entraînant de mauvaises prédictions sont à éviter : le sur-apprentissage et le sous-apprentissage (Figure 3). On parle de sur-apprentissage lorsque la méthode prend en compte le bruit des données (c'est-à-dire les individus trop atypiques) : le modèle prédictif est excessivement complexe et prédit « trop bien ». On parle de sous-apprentissage lorsque la méthode n'a pas assimilé la structure sous-jacente des données : le modèle est trop simple et ne prédit « pas assez bien ».

FINANCÉ PAR ANR
Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

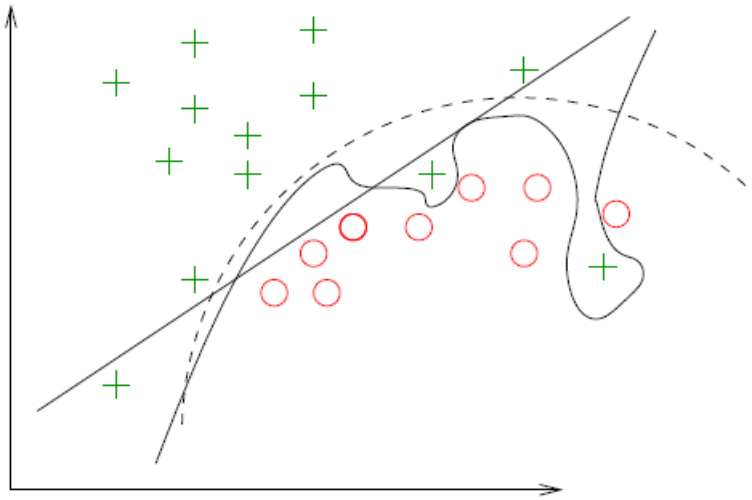


Figure 3 : Illustration du sur- et du sous-apprentissage.
Le trait plein linéaire représente un modèle en sous-apprentissage, le trait plein non-linéaire représente un modèle en sur-apprentissage, le trait pointillé non-linéaire représente un modèle au meilleur compromis.

Ce sont ces méthodes de classification supervisée que l'on se propose d'utiliser pour déterminer automatiquement le type d'un îlot urbain à partir des indicateurs urbains qui lui sont propres. Pour obtenir le meilleur taux de prédiction de notre typologie, ce rapport expose la comparaison entre 6 méthodes de classification supervisée : arbre de décision, forêt aléatoire, analyse factorielle discriminante, analyse discriminante quadratique, k plus proches voisins et machine à vecteurs supports. Ces méthodes ayant toutes été développées sous le logiciel R, toutes nos analyses ont été conduites avec le logiciel R version 3.1.3 (R Core Team, 2015).

Arbre de décision (CART = Classification And Regression Tree)

Un arbre de décision se construit par partitionnement récursif binaire des décisions selon des règles sur les variables explicatives. Un arbre de décision (figure 4) commence par une racine qui contient tous les individus. Cette racine se subdivise ensuite en deux branches à partir d'un nœud que l'on nomme nœud père. Au bout de chacune des branches apparaît alors un nœud fils qui peut à son tour se subdiviser en deux branches et donc devenir nœud père. Si un nœud ne se subdivise pas, il est appelé nœud terminal ou feuille car situé à l'extrémité d'une branche. A chaque nœud non terminal correspond une règle logique de la forme « SI...ALORS...SINON ». A chaque feuille correspond une classe. Un arbre de décision est donc une suite de règles logiques de la forme « SI...ALORS...SINON » permettant de dissocier les individus en classes.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

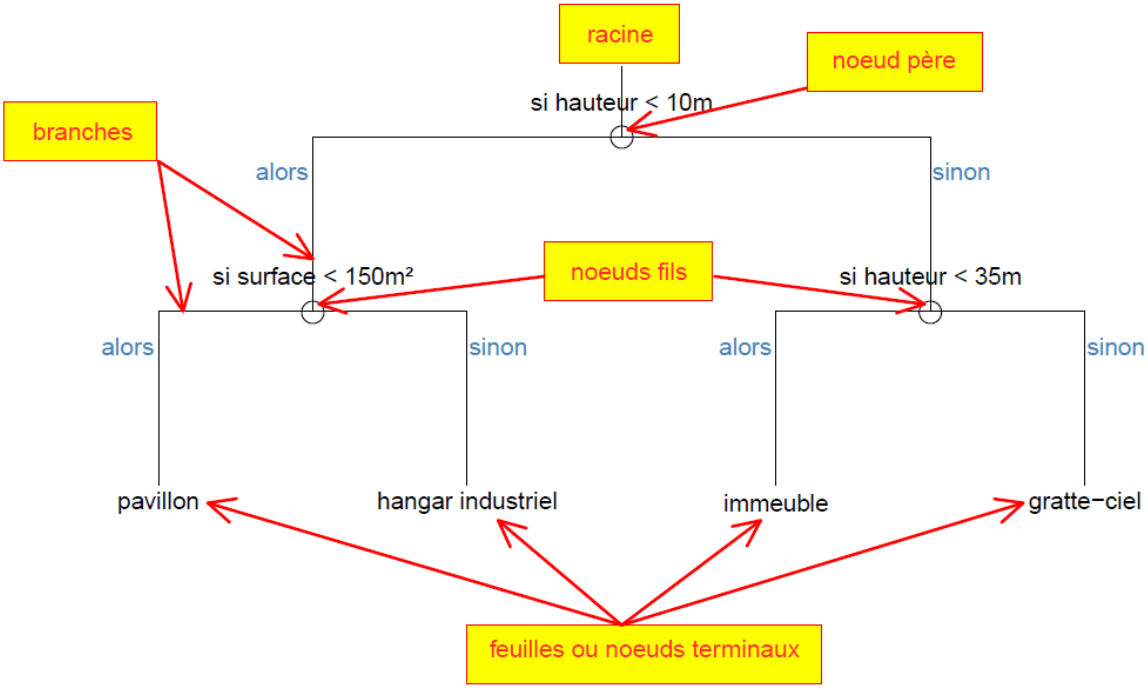


Figure 4 : Représentation graphique d'un arbre de décision

On parle d'arbre de régression lorsque la variable expliquée est de type numérique et que l'on cherche à prédire une valeur la plus proche proche possible de la vraie valeur. On parle d'arbre de classification lorsque la variable expliquée est de type facteur (comme c'est le cas pour notre typologie urbaine) ; on cherche alors, à chaque étape de partition, à réduire l'impureté (hétérogénéité d'individus appartenant à différentes classes) totale des deux nœuds fils par rapport au nœud père.

Il existe plusieurs fonctions dans R qui permettent de construire un arbre de classification, notamment la fonction **tree()** du package « tree » (Ripley, 2014) qui cherche l'arbre optimal (c'est à dire l'arbre au plus faible taux d'erreur de prédiction) et la fonction **rpart()** du package « rpart » (Therneau *et al.*, 2015) qui cherche l'arbre le plus court permettant de dissocier tous les différents groupes. Etant donné que nous cherchons la méthode la plus fiable possible, nous nous focaliserons donc sur la fonction **tree()**.

Cette fonction **tree()** crée un arbre de décisions binaires en cherchant à chaque nœud quelle variable et quelle partition permettent de maximiser la réduction d'impureté. Pour les variables numériques, une partition correspond à une valeur seuil ; pour les variables de type facteur, elle correspond à la division en 2 groupes de valeurs des facteurs.

Particularités de la méthode :
(1) Classification facile à interpréter, il est donc très facile d'y incorporer des règles expertes suivant le domaine traité.

Forêt aléatoire (Random Forest)

Une forêt aléatoire est un ensemble d'arbres de décision construits chacun avec un sous-échantillon de l'échantillon d'apprentissage. Pour chaque arbre, la sélection de la meilleure partition à un nœud est réalisée à

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

partir d'un sous-ensemble de variables tirées au hasard dans l'ensemble des variables de départ. Les variables de ce sous-ensemble sont nommées « variables présélectionnées ». On utilise ensuite l'ensemble des arbres de décision ainsi produits pour réaliser la prédiction ; soit avec un vote à la majorité pour de la classification (variable de type facteur comme notre typologie urbaine), soit avec une moyenne des valeurs pour de la régression (variable de type numérique). La fonction R correspondant à cette méthode est **randomForest()** du package « randomForest » (Liaw & Wiener, 2002). L'argument **mtry** permet d'indiquer le nombre de variables présélectionnées et l'argument **ntree** le nombre d'arbres constituant la forêt.

Particularités de la méthode :
(1) Permet de déterminer les variables qui ont le plus d'importance en tant que variable discriminante
(2) On peut combiner plusieurs forêts après coup et donc faire les calculs en parallèle ; autrement dit, au lieu de calculer une forêt de 500 arbres, on peut calculer en parallèle 5 forêts de 100 arbres.

Analyse factorielle discriminante (AFD) ou analyse discriminante linéaire (LDA = Linear Discriminante Analysis)

Cette analyse revient à faire une ACP sur le nuage des centres de gravités des groupes pondérés par leurs effectifs. L'AFD recherche des fonctions discriminantes qui permettent de différencier les classes existantes et d'affecter un nouvel individu à une classe en y associant une probabilité. Les fonctions discriminantes sont des combinaisons linéaires des variables d'entrée, d'où le nom LDA. La première fonction discriminante maximise la variance inter-classe tout en minimisant la variance intra-classe. Les autres fonctions discriminantes cherchent ensuite à discriminer au mieux les groupes. Ainsi plus on aura de fonctions, meilleure sera la discrimination. L'AFD est l'analyse d'un nuage de points caractérisée par la distance de Mahalanobis qui correspond à la distance entre les observations et les centres des groupes. Cela signifie que pour classer un nouvel individu de classe inconnue, on regarde le groupe pour lequel cette distance est minimale.

Pour réaliser ce type d'analyse, deux fonctions sont utiles dans R : les fonctions **lda()** du package « MASS » (Venables & Ripley, 2002) et **discrimin()** du package « ade4 » (Dray & Dufour, 2007).
La fonction **lda()** permet de déterminer les fonctions discriminantes linéaires par apprentissage puis de réaliser des prédictions sur de nouveaux individus.
La fonction **discrimin()** permet quant à elle de visualiser des résultats graphiques de l'AFD. Cela est particulièrement intéressant pour identifier les variables les plus discriminantes.

Particularités de la méthode :
(1) Visualiser graphiquement la répartition des classes suivant les variables.

Analyse discriminante quadratique (QDA = Quadratic Discriminant Analysis)

La méthode QDA fonctionne comme l'AFD, à la seule différence qu'elle recherche les meilleures fonctions discriminantes quadratiques, c'est-à-dire non-linéaires. Elle est implémentée dans R avec la fonction **qda()** du package « MASS » (Venables & Ripley, 2002).

Particularités de la méthode :
(1) Permet de bien classer les groupes mal classés par l'AFD car dissociables selon des fonctions non-linéaires.

K plus proches voisins (KNN = K Nearest Neighbor)

Cette méthode de classification supervisée est singulière car elle ne dispose pas de phase d'apprentissage. En effet, pour prédire la classe d'un nouvel individu, cette méthode se contente d'observer la classe des k individus voisins les plus proches du nouvel individu. Elle choisie alors pour le nouvel individu la classe majoritairement observée chez ses k voisins les plus proches. La distance entre voisins est estimée selon une mesure similarité ; c'est-à-dire que plus les caractérisques des différentes variables sont similaires entre 2 individus, pluse ces deux

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

individus seront proches. Le choix du k (nombre des plus proches voisins à observer) est important : un k trop faible va entraîné une méthode trop sensible en sur-apprentissage. Plus le k sera grand, moins la classification sera sensible au bruit et meilleure sera l'affection, mais plus le temps de calculs sera lourd. La méthode est implémentée dans R avec la fonction **knn()** du package « FNN » (Beygelzimer *et al.*, 2013).

Particularités de la méthode :
(1) Pas de phase d'apprentissage

Machine à vecteurs supports (SVM = Support Vector Machine)

Une SVM est un algorithme d'apprentissage cherchant l'hyperplan (frontière de décision) de marge optimale qui sépare au mieux les données en groupes (Figure 5). La marge correspond à la distance du point le plus proche de l'hyperplan. L'individu positionné en ce point est qualifié de vecteur support. La marge optimale est déterminée en maximisant la marge, ce qui signifie que l'hyperplan doit se trouver le plus loin possible de tous les points représentant les individus, ce qui limite le sur-apprentissage. A un hyperplan correspond une fonction linéaire. Lorsque l'hyperplan ne peut suivre une fonction linéaire, plutôt que de choisir une transformation non-linéaire, la SVM projette les données dans un espace de grande dimension où les données sont linéairement séparables (Figure 6). Cette projection est une transformation basée sur une(des) fonction(s) noyau qui calcule(nt) un produit scalaire. Cette fonction peut-être linéaire, polynomiale ou gaussienne. La méthode peut s'utiliser sous R avec la fonction **svm()** du package « e1071 » (Meyer *et al.*, 2014).

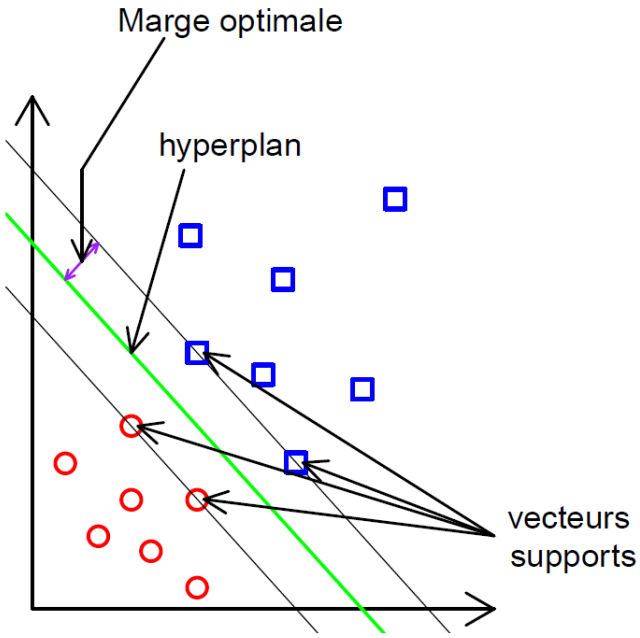


Figure 5 : Hyperplan et marge optimale

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

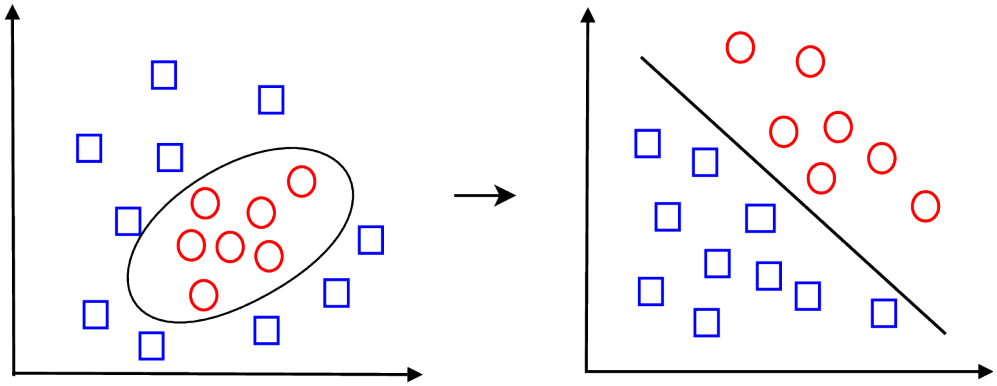


Figure 6 : Projection des données d'entrée dans un espace où elles sont linéairement séparable.

Particularités de la méthode :
(1) Limite au maximum les risques de sur-apprentissage.

B.2. Indicateurs urbains

Les indicateurs urbains fournis par l'IRSTV ont été calculés sous OrbisGIS. Certains indicateurs ont été calculés à l'échelle du bâtiment, d'autres à l'échelle du bloc de bâtiments et enfin d'autres à l'échelle de l'USR. Ces indicateurs, au nombre de 84, vont servir de variables prédictives à notre méthode de classification. Leurs calculs sont décrits dans le tableau 5.

Tableau 5 : Description des 84 indicateurs urbains utilisés en tant que variables prédictives pour la classification. Les variables dont le code commence par « i » sont calculées à l'échelle du bâtiment, celles dont le code commence par « b » à l'échelle du bloc de bâtiments, et celles dont le code commence par « u » à l'échelle de l'USR.

Code	Description
i_H_ORIGIN	Hauteur du bâtiment
i_ROOF	Hauteur de toit = $i\_ZMAX - i\_ZMIN$
i_INHAB	Nombre d'habitants par bâtiment : Nombre d'habitants des données INSEE réparties proportionnellement à la surface de plancher (i_FLOOR) des bâtiments sans nature particulière (PAI_NATURE = vide) qui sont donc potentiellement des habitations.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

i_H	Hauteur corrigée des bâtiments (hauteur utilisée pour les calculs des autres variables) :
	si i_H_ORIGIN = 0,
	alors{
	si $\frac{H\_STD_{USR}}{H\_MEAN_{USR}} < 0{,}5$
	alors { HAUTEUR_CORIGEE = $\lvert arrondi(H\_MEAN_{USR} - H\_STD_{USR})\rvert$ }
	sinon {HAUTEUR_CORIGEE = $arrondi(H\_MEAN_{USR})$ }
	}
	Sinon{HAUTEUR_CORIGEE = i_H_ORIGIN}
	
	Si HAUTEUR_CORIGEE = 0 (cas des USR où tous les bâtiments ont une hauteur égale à 0)
	alors {HAUTEUR_CORRIGEE = 3}

i_LEVELS	Nombre de niveaux d'un bâtiment calculé selon la nature du bâtiment.
	Pour le BATI_INDIFFERENCIE :
	si $HAUTEUR\_CORIGEE \geq 3$
	alors {
	NB_NIV= $arrondi(\frac{HAUTEUR\_CORIGEE - 4}{3}) + 1$
	}
	sinon{NB_NIV=1 }
	Pour les bâtiments des couches BATI_INDUSTRIEL et BATI_REMARQUABLE, se référer aux documents "etude_niveau_bati_industriel.odt" et "etude_niveau_bati_remarquable.odt"

i_AREA	Surface au sol du bâtiment (surface du polygone de la BD Topo)
i_FLOOR	Surface de plancher du bâtiment = $i\_AREA * i\_LEVELS$
i_VOL	Volume du bâtiment = $i\_H * i\_AREA$
i_COMP_B	Coefficient de Compacité brute du bâtiment = $\frac{\sum(Surfaces\,extérieures)}{Volume^{2/3}}$ = $\frac{Périmètre * Hauteur + Area}{(Hauteur * Area)^{2/3}}$ = $\frac{i\_PERI * i\_H + i\_AREA}{(i\_H * i\_AREA)^{2/3}}$
i_COMP_N	Coefficient de Compacité nette du bâtiment = $\frac{\sum(Surfaces\,extérieures\,non\,contigües)}{Volume^{2/3}}$ = $\frac{i\_FREE\_EXT\_AREA}{i\_VOL^{2/3}}$
i_COMP	Indice de compacité de Gravelius du bâtiment = $\frac{Périmètre}{2\sqrt{\pi * Area}}$ = $\frac{i\_PERI}{2\sqrt{\pi * i\_AREA}}$
i_FORM	Facteur de forme du bâtiment = $\frac{Area}{Périmètre^2}$ = $\frac{i\_AREA}{i\_PERI^2}$
i_CONC	Concavité du bâtiment = $\frac{Area}{Surface\,de\,l'enveloppe\,convexe}$
i_DIR	Direction principale du bâtiment en degré (Nord-Sud = 0° ; Est-Ouest = 90°)
i_PERI	Périmètre du bâtiment
i_WALL_A	Surface totale de façade du bâtiment = $Périmètre * Hauteur$ = $i\_PERI * i\_H$

FINANCÉ PAR ANR

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

i_PWALL_L	Longueur totale de murs mitoyens du bâtiment = $\sum(\text{Longueur mur mitoyen})$
i_PWALL_A	Surface totale de murs mitoyens du bâtiment = $\sum(\text{Longueur mur mitoyen} * min(\textit{Hauteurs des 2 bâtiments mitoyens}))$
i_NB_NEI	Nombre de bâtiments mitoyens
i_FWALL_L	Longueur totale de murs non mitoyens du bâtiment = $\textit{Périmètre} - \textit{Longueur totale de murs mitoyens} = \text{i\_PERI} - \text{i\_PWALL\_L}$
i_FREE_EXT_AREA	Surface d'enveloppe extérieure non contigüe d'un bâtiment = $\text{i\_WALL\_A} - \text{i\_PWALL\_A} + \text{i\_AREA}$
i_CONTIG	Contiguïté du bâtiment = $\frac{\textit{Surface totale de murs mitoyens}}{\textit{Surface totale de façade}} = \frac{\text{i\_PWALL\_A}}{\text{i\_WALL\_A}}$
i_PASSIV_VOL	Ratio de volume passif du bâtiment = $\frac{\textit{Surface intérieure sur une largeur de 6 m depuis les façades non mitoyennes}}{\text{i\_AREA}}$
i_FRACTAL	Dimension fractale du bâtiment = $\frac{2*\log(\textit{Périmètre})}{\log(\textit{Area})} = \frac{2*\log(\text{i\_PERI})}{\log(\text{i\_AREA})}$
i_DIST_MIN	Distance minimale avec le bâtiment le plus proche dans l'USR (distance nulle attribuée à un bâtiment seul dans une USR)
i_DIST_MEAN	Distance moyenne avec les autres bâtiments de l'USR = $\frac{\sum(\textit{Distance})}{n}$ (distance nulle attribuée à un bâtiment seul dans une USR)
i_DIST_MAX	Distance minimale avec le bâtiment le plus éloigné dans l'USR (distance nulle attribuée à un bâtiment seul dans une USR)
i_DIST_STD	Écart-type de la distance moyenne avec les autres bâtiments de l'USR = $\sqrt{\frac{\sum(\textit{Distance} - \bar{\textit{Distance}})^2}{n}}$
i_NB_POINTS	Nombre de sommets du polygone du bâtiment
i_L_TOT	Périmètre extérieur du bâtiment (longueur de murs extérieurs totale sans considérer les cours intérieures)
i_L_CVX	Périmètre de la surface convexe du bâtiment
i_L_3M	Longueur de murs du bâtiment situés à 3m ou moins de la route
i_LRATIO_3M	Pourcentage de longueur de murs du bâtiment situés à 3m ou moins de la route en considérant le périmètre extérieur du bâtiment = $\frac{\text{i\_L\_3M} * 100}{\text{i\_L\_TOT}}$
i_LRATIO_CVX	Pourcentage de longueur de murs du bâtiment situés à 3m ou moins de la route en considérant le périmètre de la surface convexe du bâtiment = $\frac{\text{i\_L\_3M} * 100}{\text{i\_L\_CVX}}$
b_AREA	Surface au sol du bloc de bâtiments = $\sum \text{i\_AREA}$
b_FLOOR	Surface de plancher du bloc de bâtiments = $\sum \text{i\_FLOOR}$

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

b_VOL	Volume du bloc de bâtiments = $\sum$ i_VOL
b_H_MEAN	Hauteur moyenne pondérée des bâtiments du bloc = $\frac{\sum (\text{i\_AREA} * \text{i\_H})}{\sum \text{i\_AREA}}$
b_H_STD	Écart-type de la hauteur des bâtiments du bloc = $\sqrt{\frac{\sum (\text{i\_H} - \bar{\text{i\_H}})^2}{n}}$
b_COMP_N	Coefficient de compacité nette du bloc de bâtiments = $\frac{\sum \text{i\_FREE\_EXT\_AREA}}{(\sum \text{i\_VOL})^{2/3}}$
b_HOLES_A	Surface des cours du bloc de bâtiments
b_HOLES_P	Pourcentage de surface des cours d'un bloc de bâtiments = $\frac{Surface_{cours}}{\text{b\_AREA} + Surface_{cours}} * 100$
b_DIR	Direction principale du bloc de bâtiments en degré (Nord-Sud = 0° ; Est-Ouest = 90°)
u_VEG_A	Surface de végétation sur l'USR (BD Topo)
u_VEG_NBPAI	Nombre de PAI (Point d'Activité et d'Intérêt) nature dans l'USR
u_ROAD_A	Surface de réseau routier dans l'USR
u_ROAD_L	Longueur de réseau routier dans l'USR (ligne de la BD Topo)
u_ROAD_NBPAI	Nombre de PAI transport dans l'USR
u_SIDEWALK_L	Périmètre de l'îlot obtenu après union des parcelles contigues
u_WATER_A	Surface des éléments hydrographiques dans l'USR : réservoirs et surfaces en eau
u_WATER_L	Longueur de réseau hydrographique dans l'USR (BD Topo)
u_INHAB	Nombre d'habitants par USR : Nombre d'habitants des données INSEE réparti proportionnellement à la surface de plancher (i_FLOOR) des bâtiments d'une USR sans nature particulière (PAI_NATURE = vide) qui sont donc potentiellement des habitations.
u_HOUSE	Nombre de ménages par USR : Nombre de ménages des données INSEE réparti proportionnellement à la surface de plancher (i_FLOOR) des bâtiments d'une USR sans nature particulière (PAI_NATURE = vide) qui sont donc potentiellement des habitations.
u_COL_HOUSE	Nombre de ménages logeant dans des habitats collectifs par USR : Nombre de ménages logeant dans des habitats collectifs des données INSEE réparti proportionnellement à la surface de plancher (i_FLOOR) des bâtiments d'une USR sans nature particulière (PAI_NATURE = vide) qui sont donc potentiellement des habitations.
u_HOUSE_A	Surface totale des habitations par USR : Surface cumulée des habitations des données INSEE répartie proportionnellement à la surface de plancher (i_FLOOR) des bâtiments d'une USR sans nature particulière (PAI_NATURE = vide) qui sont donc potentiellement des habitations.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

u_COL_HOUSE_A	Surface totale des habitats collectifs par USR = $\frac{u\_COL\_HOUSE}{u\_HOUSE} * u\_HOUSE\_A$
u_FLOOR	Surface de plancher totale des bâtiments de l'USR = $\sum (i\_FLOOR)$
u_COS	Coefficient d'occupation des sols d'une USR = $\frac{Surface\ de\ plancher\ totale}{Surface\ de\ l'USR} = \frac{u\_FLOOR}{u\_AREA}$
u_COMP_NWMEAN	Compacité nette moyenne des bâtiments de l'USR = $\frac{\sum (i\_COMP\_N)}{n}$
u_COMP_WMEAN	Compacité nette moyenne pondérée des bâtiments de l'USR = $\frac{\sum (i\_AREA * i\_COMP\_N)}{\sum i\_AREA}$
u_CONTIG_MEAN	Contiguïté moyenne pondérée des bâtiments de l'USR = $\frac{\sum (i\_AREA * i\_CONTIG)}{\sum i\_AREA}$
u_CONTIG_STD	Écart-type de la contiguïté des bâtiments de l'USR = $\sqrt{\frac{\sum (i\_CONTIG - \bar{i\_CONTIG})^2}{n}}$
u_DIR_STD	Écart-type de la direction principale des bâtiments de l'USR = $\sqrt{\frac{\sum (i\_DIR - \bar{i\_DIR})^2}{n}}$
u_H_MEAN	Hauteur moyenne pondérée des bâtiments de l'USR = $\frac{\sum (i\_AREA * i\_H)}{\sum i\_AREA}$
u_H_STD	Écart-type de la hauteur des bâtiments de l'USR = $\sqrt{\frac{\sum (i\_H - \bar{i\_H})^2}{n}}$
u_PASSIV_VOL_MEAN	Ratio de volume passif moyen pondéré par la surface de plancher des bâtiments de l'USR = $\frac{\sum (i\_FLOOR * i\_PASSIV\_VOL)}{\sum i\_FLOOR}$
u_AREA	Surface totale des bâtiments de l'USR = $\sum i\_AREA$
u_VOL	Volume total des bâtiments de l'USR = $\sum i\_VOL$
u_VOL_MEAN	Volume moyen des bâtiments de l'USR = $\frac{\sum (i\_VOL)}{n}$
u_NB_BUILD	Nombre de bâtiments de l'USR
u_DIST_MIN_MEAN	Distance minimale entre bâtiments moyenne des bâtiments de l'USR = $\frac{\sum (i\_DIST\_MIN)}{n}$

FINANCÉ PAR ANR

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

u_DIST_MEAN_MEAN	Distance moyenne entre bâtiments des bâtiments de l'USR = $\frac{\sum(\text{i\_DIST\_MEAN})}{n}$
u_DIST_MEAN_STD	Écart-type des distances moyennes entre bâtiments des bâtiments de l'USR = $\sqrt{\frac{\sum(\text{i\_DIST\_MEAN}-\overline{\text{i\_DIST\_MEAN}})^2}{n}}$
u_bHOLES_A_MEAN	Surface de cours moyenne pondérée des blocs de l'USR = $\frac{\sum(\text{b\_AREA}*\text{b\_HOLES\_A})}{\sum \text{b\_AREA}}$
u_bH_STD_MEAN	Écart-type de la hauteur moyen pondéré des blocs de l'USR = $\frac{\sum(\text{b\_AREA}*\text{b\_H\_STD})}{\sum \text{b\_AREA}}$
u_bCOMP_NWMEAN	Compacité nette moyenne des blocs de l'USR = $\frac{\sum(\text{b\_COMP\_N})}{n_{bloc}}$
u_bCOMP_WMEAN	Compacité nette moyenne pondérée des blocs de l'USR = $\frac{\sum(\text{b\_AREA}*\text{b\_COMP\_N})}{\sum \text{b\_AREA}}$
u_bCOMP_STD	Écart-type de la compacité nette des blocs de l'USR = $\sqrt{\frac{\sum(\text{b\_COMP\_N}-\overline{\text{b\_COMP\_N}})^2}{n_{bloc}}}$
u_DIST_CENTER	Distance entre le centroïde de l'USR et la mairie de la ville à laquelle l'USR appartient.
u_BUILD_DENS	Densité surfacique des bâtiments de l'USR= $\frac{\sum \text{i\_AREA}}{Area_{USR}}$
u_WATER_DENS	Densité surfacique du réseau hydrographique de l'USR= $\frac{\text{u\_WATER\_A}}{Area_{USR}}$
u_VEG_DENS	Densité surfacique de la végétation de l'USR= $\frac{\text{u\_VEG\_A}}{Area_{USR}}$
u_ROAD_DENS	Densité surfacique du réseau routier de l'USR= $\frac{\text{u\_ROAD\_A}}{Area_{USR}}$
u_FWALL_A	Surface de façade extérieure (non contigue) totale des bâtiments de l'USR = $\sum(\text{i\_WALL\_A}-\text{i\_PWALL\_A})$

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

B.3. Acquisition des données typologiques d'apprentissage

Protocole d'identification typologique de terrain

Comme expliqué dans la partie B.1., pour pouvoir utiliser des méthodes supervisées, il est indispensable de connaître à la fois les variables et la typologie de certains individus pour la phase d'apprentissage. Les variables correspondent aux indicateurs urbains calculés par l'IRSTV et sont donc disponibles. En revanche, bien que nous connaissions les différents types de bâtiments urbains présents en France, soit 10 types (c.f. partie A.3.), nous ne disposions pas de données réelles. Il a donc fallu établir un protocole d'identification typologique de terrain (Figure 7) pour pouvoir en obtenir. Ce protocole sous forme d'arbre décisionnel s'est créé et optimisé par boucle de rétroaction entre les résultats d'enquête de la tâche 1.2 (cf A.3) et des observations de terrain menées au travers d'images satellites et de représentations tridimensionnelles issues du logiciel Google Earth.

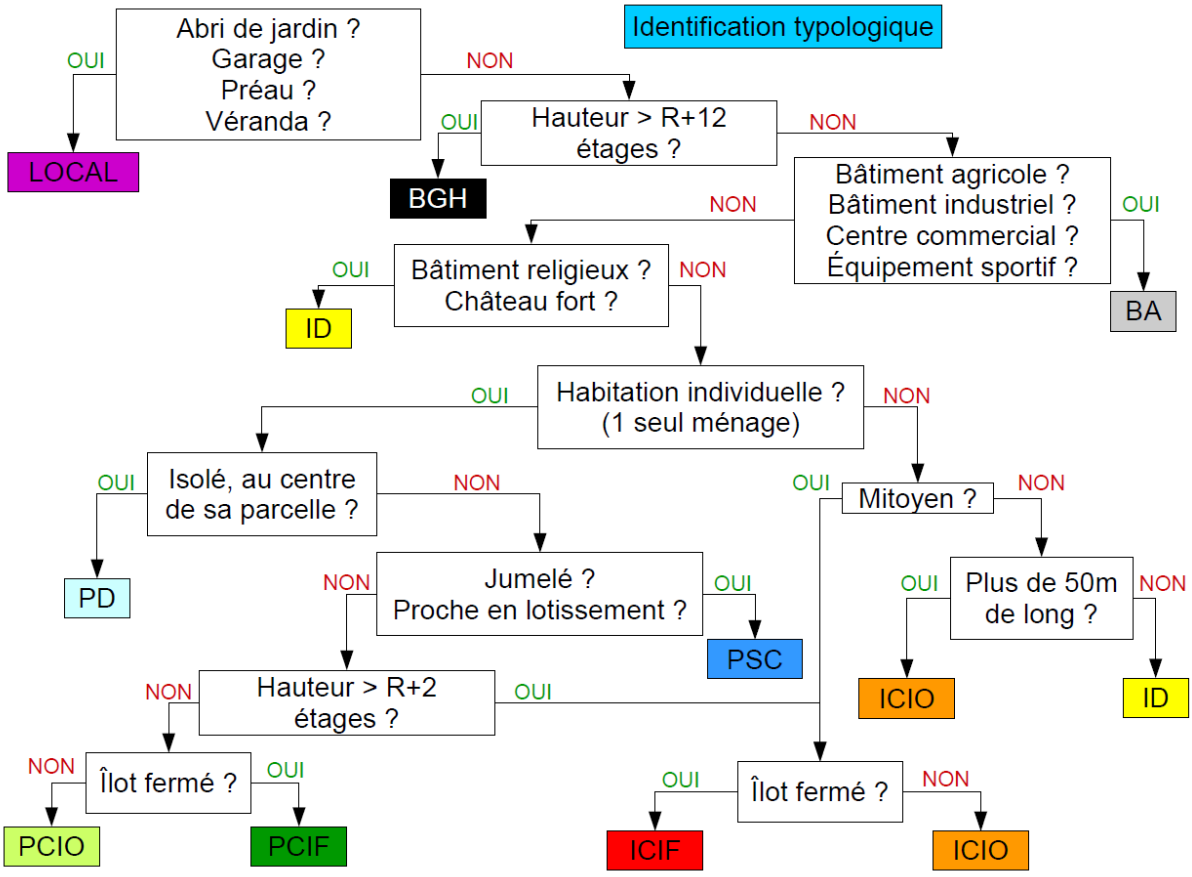


Figure 7 : Protocole d'identification d'un bâtiment selon la typologie élaborée dans le cadre du projet MapUCE. BA= bâtiment d'activité ; BGH= bâtiment de grande hauteur ; ICIF= immeuble continu sur îlot fermé ; ICIO= immeuble continu sur îlot ouvert ; ID= immeuble discontinu ; LOCAL= local annexe ; PCIF= pavillon continu sur îlot fermé ; PCIO= pavillon continu sur îlot ouvert ; PD= pavillon discontinu ; PSC= pavillon semi-continu.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Choix des cas d'étude

Pour pouvoir mettre en place une classification typologique adaptée à toute la France, il est nécessaire d'inclure plusieurs cas d'étude répartis sur l'ensemble du territoire français. Notre choix s'est porté sur sept cas d'étude : Annecy, La Rochelle, Avignon, Toulouse, Nantes, Paris et Mulhouse. Annecy a été choisie pour son architecture adaptée à un environnement montagnard ; La Rochelle pour sa situation estuarienne et la forte présence de matériaux calcaire ; Avignon pour son architecture « provençale » ; Toulouse pour la dominance des constructions en brique ; Nantes pour l'influence de l'architecture bretonne avec murs en pierre (granit) et toitures en ardoises ; Mulhouse pour sa localisation frontalière dont l'architecture est potentiellement influencée par les pays limitrophes de l'EST comme l'Allemagne ; et enfin Paris, composée des départements de Paris (75) et des Hauts-de-Seine (92), zone la plus dense de France en bâtiments et en population.

Mise en application du protocole

Le protocole d'identification des typologies ci-avant présenté a été appliqué en utilisant le logiciel Google Earth. Pour chacun des sept cas d'étude, lorsque les typologies réelles le permettaient, un minimum de huit USR typiques par type a été sélectionné. Tous les bâtiments contenus dans ces USR ont été identifiés et donc classés dans l'un des dix types. Nous disposions de données obtenues avec le logiciel OrbisGIS comprenant pour chaque bâtiment la forme du polygone et les indicateurs calculés. Les types identifiés ont ainsi été fusionés aux données via le logiciel ArcGIS.

B.4. Changement d'échelle de classification : du bâtiment à l'USR

Attribuer un type à un bâtiment reste sommaire : selon ses caractéristiques morphologiques, les caractéristiques morphologiques du bloc auquel il appartient ainsi que les indicateurs environnementaux de l'USR auquel il appartient, un bâtiment sera classé dans un des types définis en A.3. Attribuer un type à une USR est un exercice légèrement plus complexe. En effet, l'hétérogénéité des types de bâtiments au sein d'un îlot (tableau 6) ainsi que l'hétérogénéité des types d'îlots au sein d'une USR (se reporter au tableau 8) posent la question suivante : à quel type de bâtiment présent sur l'USR doit être attribuée l'USR ?

Tableau 6 : Exemples d'îlots hétérogènes en types de bâtiments

Typologies	Photographie de la réalité	Représentation sous SIG
pcif (vert) icif (rouge) local (violet)		

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

pd (bleu clair) psc (bleu foncé) local (violet)		
id (jaune) icio (orange)		

icif= immeuble continu sur îlot fermé ; icio= immeuble continu sur îlot ouvert ; id= immeuble discontinu ; local= local annexe ; pcif= pavillon continu sur îlot fermé ; pd= pavillon discontinu ; psc= pavillon semi-continu.

Plutôt que de donner un type unique à une USR qui peut en contenir jusqu'à dix différents, on se propose de déterminer les proportions de chaque type de bâtiment et d'exploiter les types majoritaire et secondaire d'une USR. Le type majoritaire se défini comme étant le type de bâtiment le plus dominant dans l'USR, le type secondaire comme étant le type de bâtiment le deuxième plus dominant dans l'USR. Cette dominance est estimée à travers le pourcentage de surface de plancher de chaque type de bâtiment dans l'USR. Ainsi, le type le plus dominant d'une USR, autrement dit le type majoritaire d'une USR, est celui dont la somme des surfaces de plancher des bâtiments de ce type contenus dans l'USR est la plus élevée. Le choix de la surface de plancher s'est décidé afin de minimiser la dominance des bâtiments ayant peu d'influence sur le climat et la consommation énergétique. Prenons pour exemple un immeuble d'habitation de deux étages et un grand hangar de stockage ne comportant aucun étage. Ces deux bâtiments ont la même hauteur, la même emprise au sol et donc le même volume. Néanmoins, la consommation énergétique et l'émission de chaleur du hangar sont beaucoup moins importantes que celles de l'immeuble d'habitation. Il faut donc utiliser la surface de plancher, trois fois plus grande pour l'immeuble que pour le hangar, pour traduire cette dominance en terme de consommation énergétique et d'influence climatique.

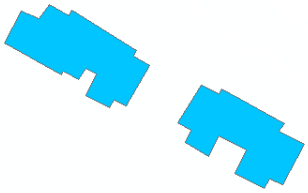
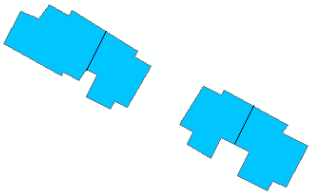
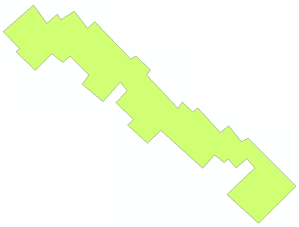
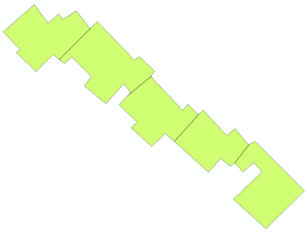
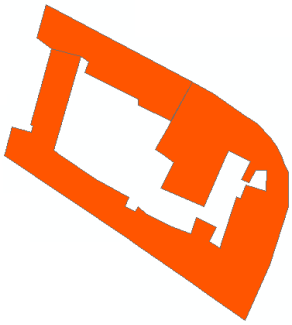
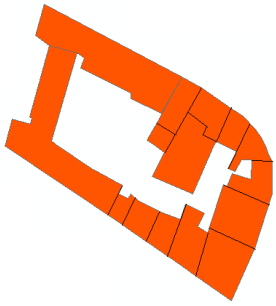
Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

C. Phase 1 : sélection des données d'apprentissage

C.1. Objectif

La typologie de bâtiments décrite en A.3. est par définition une représentation de la réalité qui écarte les bâtiments atypiques. Notre classification finale aura donc un certain pourcentage d'erreur de prédiction car incapable de prédire parfaitement les cas atypiques, ce qui est tout à fait normal. Il se trouve néanmoins que certaines données provenant de la BD TOPO ne correspondent pas à la réalité ; le cas le plus courant étant la fusion de plusieurs bâtiments en un seul. On travaille alors sur des données erronées créant, en plus des bâtiments atypiques réels, des bâtiments que l'on peut qualifier d'atypiques irréels dont voici quelques exemples tableau 7. Ainsi, les indicateurs calculés pour ces bâtiments atypiques, comme la surface d'emprise au sol ou le nombre de bâtiments mitoyens, ne seront pas représentatifs de la réalité.

Tableau 7 : Exemples de données atypiques irréelles

Typologie	Photographie de la réalité	Représentation atypique irrélle	Représentation réelle attendue
psc			
pcio			
icif			

icif= immeuble continu sur îlot fermé ; pcio= pavillon continu sur îlot ouvert ; psc= pavillon semi-continu.

Une autre source de problème pour la classification est le fait que certaines USR sont définies de trop grande taille et peuvent ainsi réunir plusieurs îlots de quartier qui auraient dus être séparés en plusieurs USR (Tableau 8

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

pas encore dessiné). Ces USR démesurées faussent les calculs de certains indicateurs, comme le nombre de bâtiments dans l'USR ou la distance minimale moyenne d'un bâtiment aux autres bâtiments de l'USR, et donc le profil morphologique des bâtiments de cette USR devient un profil atypique irréel.

Tableau 8 : Exemple d'USR de grande taille réunissant plusieurs îlots de différents types.

Photographie de la réalité	Représentation de l'USR
	

icio= immeuble continu sur îlot ouvert (orange) ; id= immeuble discontinu (jaune) ; local= local annexe (violet) ; pcio= pavillon continu sur îlot ouvert (vert clair); pd= pavillon discontinu (bleu clair) ; psc= pavillon semi-continu (bleu foncé).

Dans tous les cas, ces bâtiments atypiques irréels vont impacter les prédictions de la classification. Pour savoir si les données relatives à ces bâtiments atypiques doivent être prises en compte dans l'apprentissage de la classification, il est nécessaire d'évaluer leur impact sur la classification, ce que nous nous sommes proposés de faire dans cette phase 1.

C.2. Méthode

Incorporation de données atypiques

Pour chacun des sept cas d'étude, en plus des données typiques déjà identifiées, des USR contenant une grande hétérogénéité de types de bâtiments ainsi que des USR contenant des bâtiments fusionnés ont été identifiés selon la méthode exposée en B.3. Pour ce qui est de la quantité de données, on s'est cette fois ci basé sur l'égalité du nombre de bâtiments par type. On a ainsi essayé autant que possible d'avoir autant de bâtiments dans chacun des types, en rajoutant si besoin des bâtiments typiques. On dispose alors de deux jeux de données par cas d'étude : un jeu données typiques et un jeu de données typiques et atypiques que l'on nommera jeu de données atypiques pour plus de facilité de lecture. Le résumé de leur composition typologique est présenté dans les tableaux 9 et 10.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Tableau 9 : Composition typologique à l'échelle du bâtiment et typologique majoritaire à l'échelle de l'USR des données typiques des cas d'études d'Annecy, La Rochelle, Avignon, Toulouse, Nantes, Paris et Mulhouse.

Typologie du bâtiment	Nombre de bâtiments							
	Annecy	La Rochelle	Avignon	Toulouse	Nantes	Paris	Mulhouse	TOTAL
ba	118	138	136	186	185	216	220	1199
bgh	4	6	9	18	28	86	67	218
icif	89	189	103	79	112	202	213	987
icio	41	83	31	61	72	132	155	575
id	63	36	45	101	47	112	66	470
local	71	178	124	74	256	270	636	1609
pcif	14	190	169	19	157	145	144	838
pcio	89	110	90	42	164	138	139	772
pd	174	171	111	191	166	189	112	1114
psc	123	154	81	79	230	192	96	955
TOTAL	786	1255	899	850	1417	1682	1848	8737

Typologie majoritaire de l'USR	Nombre d'USR							
	Annecy	La Rochelle	Avignon	Toulouse	Nantes	Paris	Mulhouse	TOTAL
ba	8	8	8	8	8	8	8	56
bgh	1	3	4	7	9	8	5	37
icif	8	9	8	8	8	10	8	59
icio	9	14	8	8	9	9	8	65
id	8	8	9	8	10	9	8	60
local	0	0	0	0	0	0	0	0
pcif	3	8	8	8	8	8	8	51
pcio	8	8	8	8	8	8	8	56
pd	10	8	8	8	8	8	8	58
psc	7	9	8	8	8	8	8	56
TOTAL	62	75	69	71	76	76	69	489

Ba= bâtiment d'activité, bgh= bâtiment de grande hauteur, icif= immeuble continu sur îlot fermé, icio= immeuble continu sur îlot ouvert, id= immeuble discontinu, local= local annexe, pcif= pavillon continu sur îlot fermé, pcio= pavillon continu sur îlot ouvert, pd= pavillon discontinu, psc= pavillon semi-continu.

FINANCÉ PAR ANR

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Tableau 10 : Composition typologique à l'échelle du bâtiment et typologique majoritaire à l'échelle de l'USR des données atypiques des cas d'études d'Annecy, La Rochelle, Avignon, Toulouse, Nantes, Paris et Mulhouse.

Typologie du bâtiment	Nombre de bâtiments							
	Annecy	La Rochelle	Avignon	Toulouse	Nantes	Paris	Mulhouse	TOTAL
ba	306	363	257	310	274	330	325	2165
bgh	4	6	9	46	54	149	68	336
icif	147	224	230	240	325	417	275	1858
icio	194	219	184	221	233	353	363	1767
id	266	175	162	226	160	283	154	1426
local	142	233	148	97	376	376	664	2036
pcif	15	475	338	68	353	239	238	1726
pcio	254	616	474	109	444	257	398	2552
pd	383	373	347	390	346	323	288	2450
psc	281	367	251	152	402	294	296	2043
TOTAL	1992	3051	2400	1859	2967	3021	3069	18359

Typologie majoritaire de l'USR	Nombre d'USR							
	Annecy	La Rochelle	Avignon	Toulouse	Nantes	Paris	Mulhouse	TOTAL
ba	39	34	15	12	10	12	15	137
bgh	1	3	4	14	14	26	5	67
icif	13	15	23	50	13	30	12	156
icio	44	56	36	24	29	37	40	266
id	34	35	22	12	17	19	23	162
local	0	0	0	0	0	0	0	0
pcif	3	32	11	17	14	10	15	102
pcio	22	34	36	14	22	11	14	153
pd	26	10	13	10	9	9	10	87
psc	18	13	10	8	9	9	14	81
TOTAL	200	232	170	161	137	163	148	1211

Ba= bâtiment d'activité, bgh= bâtiment de grande hauteur, icif= immeuble continu sur îlot fermé, icio= immeuble continu sur îlot ouvert, id= immeuble discontinu, local= local annexe, pcif= pavillon continu sur îlot fermé, pcio= pavillon continu sur îlot ouvert, pd= pavillon discontinu, psc= pavillon semi-continu.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Analyse des données

Pour chacun des deux jeux de données (typiques et atypiques) de chacun des sept cas d'étude (Annecy, La Rochelle, Avignon, Toulouse, Nantes, Paris et Mulhouse), cinq méthodes de classification ont été appliquées : Tree, RandomForest, LDA, KNN et SVM. La méthode QDA n'a pas pu être appliquée, car le nombre de bâtiments par type doit être supérieur ou égal au nombre de variables utilisées +1. Disposant de 84 variables, il faut donc impérativement 85 bâtiments par type pour pouvoir utiliser cette méthode. Or d'après les tableaux 9 et 10, le nombre de bâtiments de type « bgh » ou « id » est souvent trop petit. Pour la méthode RandomForest, le nombre de variables présélectionnées **mtry** a été fixé a 2 et le nombre d'arbres **nTree** à 500. Pour la méthode KNN, le nombre de voisins **k** a été fixé à 3. Pour les méthodes LDA et SVM, la variable « u_VEG_NBPAI » a été retirée pour les jeux de données où elle était constante.
Pour estimer la qualité d'un modèle prédictif, nous avons implémenté sous R une validation 70/30. Cette validation consiste à prendre au hasard 70 % des données de chaque type pour réaliser l'apprentissage de la classification, puis à prédire les 30 % des données de chaque type restant avec le modèle de classification créé. Les pourcentages d'erreur de classification des bâtiments et des typologies majoritaires des USR sont ensuite calculés en croisant les types prédits et observés. Cette opération est répétée cent fois et une moyenne des cent pourcentages d'erreur est calculée. On obtient alors pour chacune des cinq méthodes de classification quatorze pourcentages d'erreur moyens de classification des bâtiments et quatorze pourcentages d'erreur moyens de classification des typologies majoritaires des USR ; dont à chaque fois sept valeurs (une par cas d'étude) relatives aux données typiques et sept autres relatives aux données atypiques. Un test de comparaison non paramétrique de Wilcoxon a été réalisé entre les résultats des données typiques et atypiques appariées (car appartenant aux mêmes cas d'étude) de chaque méthode de classification pour tester l'hypothèse d'une différence de pourcentages d'erreur de classification moyens entre données typiques et atypiques.

C.3. Résultats

Comme présenté dans la figure 8, l'utilisation de données atypiques a tendance à augmenter le pourcentage d'erreur de classification moyen. Cette tendance est plus ou moins prononcée suivant l'échelle et la méthode de classification. Elle est en effet nettement plus significative, d'après les tests de Wilcoxon, à l'échelle de l'USR où seuls les résultats des données typiques et atypiques de la méthode SVM ne sont pas significativement différents à 5 %. A l'échelle du bâtiment, les résultats des méthodes RandomForest, KNN et SVM ne présentent pas de différence significative entre données typiques et atypiques ; il semblerait donc que ces méthodes de classification traitent aussi bien les données typiques qu'atypiques. A l'inverse, les méthodes Tree et LDA ont des pourcentages d'erreur moyens significativement plus faibles avec les données typiques et donnent donc des résultats significativement meilleurs avec les données typiques ; leurs performances sont donc dépendantes des données. En définitive, il est nécessaire d'inclure les données atypiques dans nos données d'apprentissage pour la classification typologique, car cela influence plus ou moins les résultats prédictifs selon la méthode de classification utilisée. De plus, le modèle final doit être capable de traiter aussi bien les données typiques que les données atypiques très présentes dans les données du territoire français.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

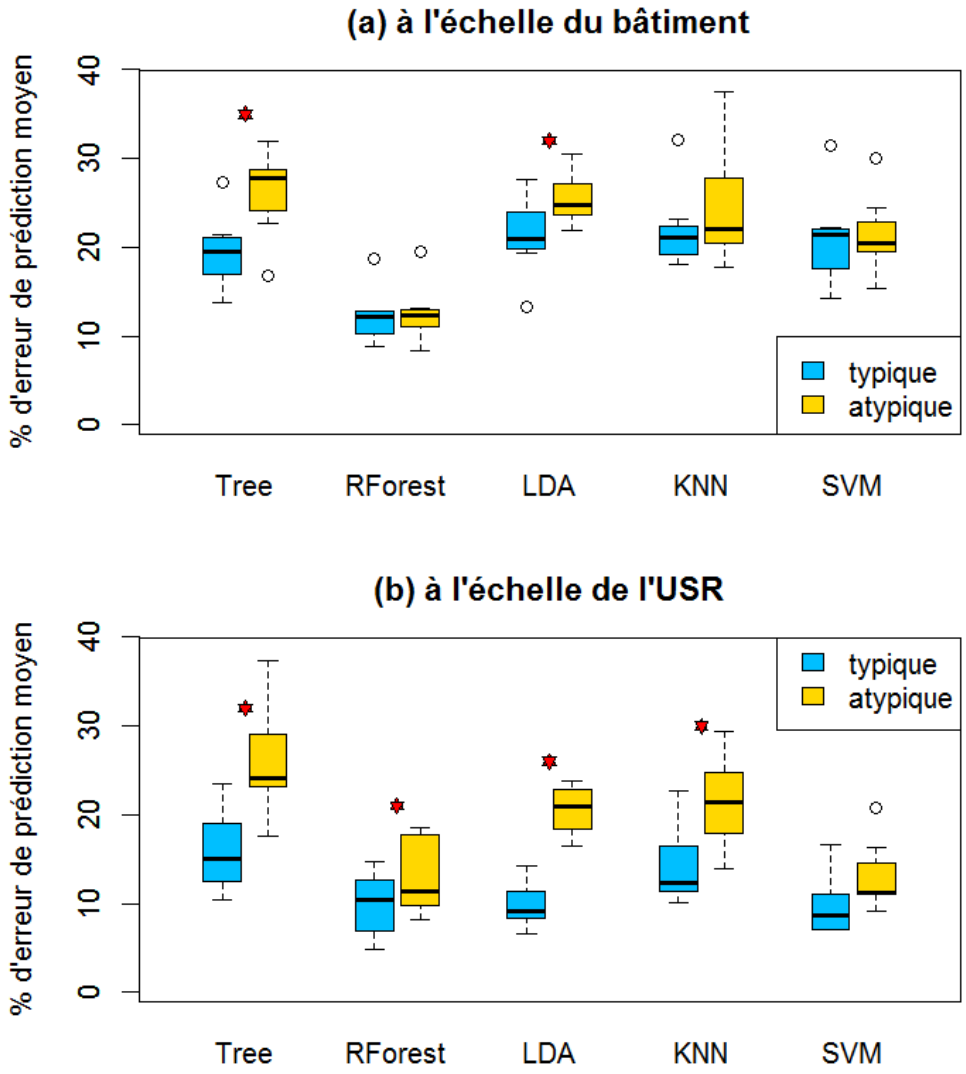


Figure 8 : Boxplots des pourcentages d'erreur de prédiction moyens obtenus (a) à l'échelle du bâtiment et (b) à l'échelle de l'USR, sur données typiques en bleu et données typiques+atypiques en jaune, pour chacune des méthodes de classification : Tree, RandomForest, LDA, KNN et SVM. Une étoile rouge indique une différence significative à 5 % (test non paramétrique de Wilcoxon) entre les pourcentages d'erreur de prédiction moyens obtenus avec des données typiques et typiques+atypiques pour une même méthode de classification.

D. Phase 2 : sélection de la méthode de classification

D.1. Objectifs

Comme cela a été démontré en phase 1, la qualité des prédictions typologiques dépend des données d'apprentissage, mais également de la méthode de classification utilisée ; chaque méthode ayant son propre système d'attribution d'un bâtiment à un type. Le but est ici de déterminer le système d'attribution le plus adapté aux données pour prédire au mieux la typologie d'un bâtiment ; autrement dit d'identifier la méthode donnant le

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

moins d'erreur de classification. Un second objectif est de vérifier l'utilité des données de chaque cas d'étude dans le modèle global.

D.2. Méthode

Pour chacun des sept cas d'étude (Annecy, La Rochelle, Avignon, Toulouse, Nantes, Paris et Mulhouse), on utilise les données atypiques de six des cas d'études pour la phase d'apprentissage, puis on prédit la typologie des données atypiques du septième cas d'étude. On calcule alors les pourcentages d'erreur de classification des bâtiments et des typologies majoritaires des USR en croisant les types prédits et observés. On parle ici d'une validation sur données exogènes, car on prédit des données n'ayant pas servies à l'apprentissage. Ces sept modèles (un par cas d'étude) sont chacun réalisés avec chacune des six méthodes de classification : Tree, RandomForest, LDA, QDA, KNN et SVM. Pour la méthode RandomForest, le nombre de variables présélectionnées **mtry** a été fixé a 2 et le nombre d'arbres **nTree** à 500. Pour la méthode KNN, le nombre de voisins **k** a été fixé pour obtenir la meilleur qualité de résultats à 3 pour les cas d'étude de La Rochelle, Avignon, Toulouse et Mulhouse, à 5 pour le cas d'étude d'Annecy, à 7 pour le cas d'étude de Nantes et à 21 pour le cas d'étude de Paris. Pour la méthode QDA, les variables « i_H », « i_PERI » et « i_FREE_EXT_AREA » ont dû être retirées pour des raisons de colinéarité avec d'autres variables au sein d'un type ; les variables « b_HOLES_A » et « b_HOLES_P » ont dû être retirées car constantes au sein d'un type ; et la variable « u_VEG_NBPAI » a été retirée pour les jeux de données de Nantes et Paris où elle était constante au sein d'un type.
Des tests de comparaison non paramétriques de Wilcoxon ont été réalisés entre les résultats appariés (car appartenant aux mêmes cas d'étude) des six méthodes de classification pour tester l'hypothèse d'une différence de pourcentages d'erreur de classification entre deux méthodes.

D.3. Résultats

A l'échelle du bâtiment, les résultats des méthodes RandomForest, KNN et QDA sont significativement différents de ceux de toutes les autres méthodes (Tableau 11). Les résultats des méthodes Tree, LDA et SVM ne sont quant à eux pas significativement différents entre eux. On peut ainsi comprendre à partir de la figure 9 que la méthode RandomForest est celle qui a significativement les plus faibles pourcentages d'erreur de prédiction, que les méthodes Tree, LDA et SVM obtiennent des pourcentages d'erreur de prédictions similaires qui sont plus élevés qu'avec la méthode RandomForest. Enfin les méthodes QDA et KNN sont les méthodes dont les pourcentages d'erreur de prédiction sont significativement les plus élevés.

A l'échelle de l'USR, les résultats de la méthode KNN sont toujours significativement différents de ceux de toutes les autres méthodes (Tableau 11) et sont toujours les moins bons (Figure 9). En revanche la distinction entre les autres méthodes est moins nette. Les pourcentages d'erreur de prédiction de la méthode RandomForest sont toujours les plus faibles (figure 9) mais ne sont pas significativement différents de ceux de la méthode SVM (Tableau 11). Les pourcentages d'erreur de prédiction des méthodes Tree, LDA et SVM ne sont pas différents entre eux et sont plus élevés que ceux de la méthode RandomForest. Enfin les pourcentages d'erreur de prédiction de la méthode QDA ne sont pas différents de ceux des méthodes Tree et SVM et sont significativement plus élevés que ceux des méthodes RandomForest et LDA (Tableau 11 et Figure 9).

Aux deux échelles, on retrouve sensiblement le même ordre des méthodes selon leurs pourcentages d'erreur de prédiction. Ainsi la méthode donnant les meilleurs résultats est RandomForest, suivie par les méthodes SVM, LDA, Tree, QDA et KNN qui est celle donnant les moins bons résultats.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Tableau 11 : Résultats des tests de comparaison de Wilcoxon entre les pourcentages d'erreurs de prédiction des six méthodes de classification : Tree, RandomForest, LDA, QDA, KNN et SVM, à l'échelle du bâtiment et à l'échelle de l'USR.

à l'échelle du bâtiment

	Rforest	KNN	LDA	QDA	SVM
KNN	0,016 *	NA	NA	NA	NA
LDA	0,016 *	0,016 *	NA	NA	NA
QDA	0,016 *	0,016 *	0,016 *	NA	NA
SVM	0,016 *	0,016 *	0,156	0,016 *	NA
tree	0,016 *	0,016 *	0,219	0,016 *	0,469

à l'échelle de l'USR

	Rforest	KNN	LDA	QDA	SVM
KNN	0,016 *	NA	NA	NA	NA
LDA	0,016 *	0,016 *	NA	NA	NA
QDA	0,016 *	0,016 *	0,031 *	NA	NA
SVM	0,297	0,016 *	0,578	0,078	NA
tree	0,022 *	0,016 *	0,295	0,109	0,297

*p-value < 0,05

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

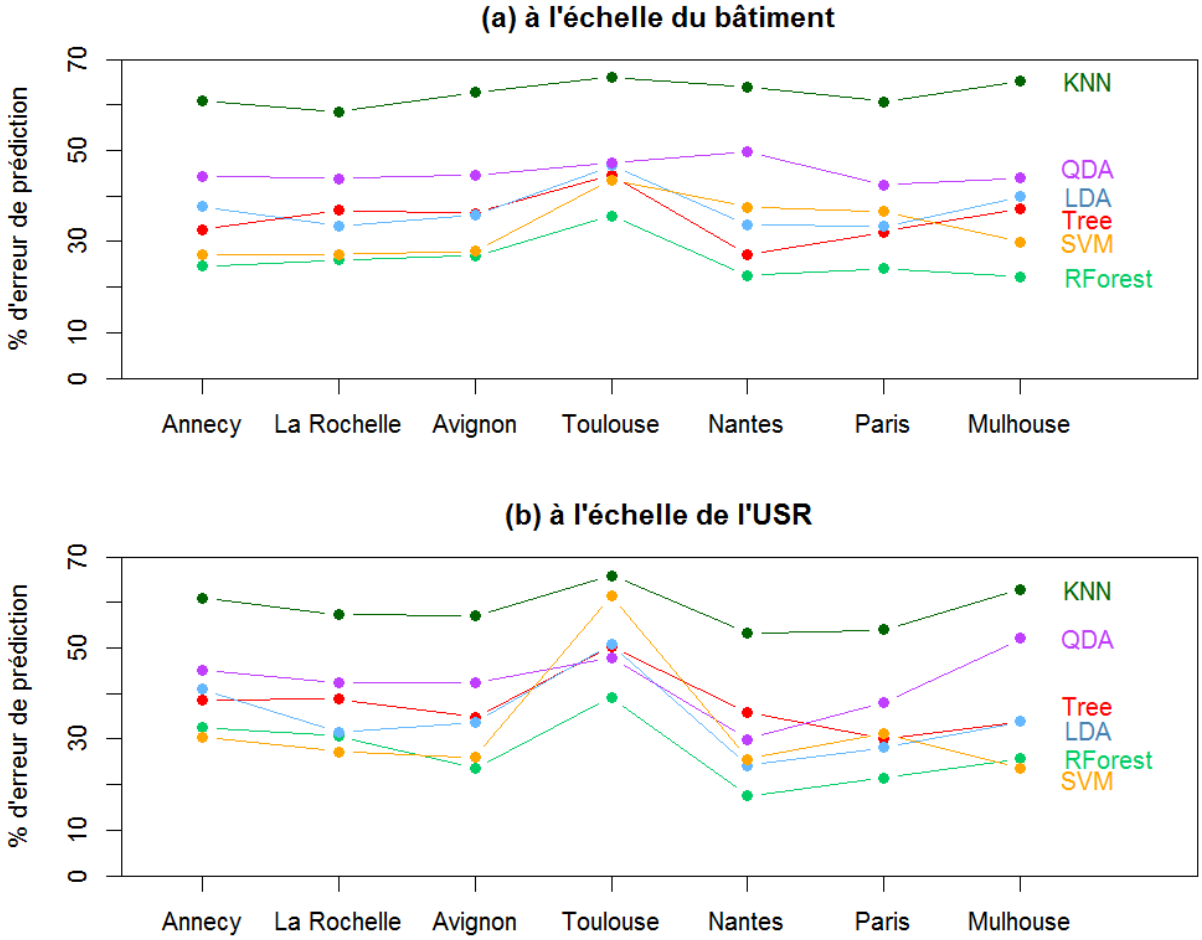


Figure 9 : Graphique des pourcentages d'erreur de prédiction (a) à l'échelle du bâtiment et (b) à l'échelle de l'USR, des cas d'étude d'Annecy, La Rochelle, Avignon, Toulouse, Nantes, Paris et Mulhouse pour six méthodes de classification : Tree, RandomForest, LDA, QDA, KNN et SVM.

Pour l'ensemble des sept cas d'étude, tous les modèles dépassent les 22 % d'erreur de prédiction à l'échelle du bâtiment et les 17 % d'erreur de prédiction à l'échelle de l'USR ; les minimums observés étant de 22,12 % pour Mulhouse à l'échelle du bâtiment et de 17,52 % pour Nantes à l'échelle de l'USR. Les données de Nantes étant les mieux prédites (avec la méthode RandomForest : 22,48 % à l'échelle du bâtiment et 17,52 % à l'echelle de l'USR) elles sont donc les moins utiles pour l'apprentissage du modèle de classification à l'échelle du territoire français. Néanmoins les pourcentages d'erreur de prédiction de Nantes sont trop élevés pour considérer les données de Nantes comme inutiles lors de l'apprentissage. Il apparaît clairement sur la figure 9 que le cas d'étude de Toulouse est le moins bien prédit car possédant les pourcentages d'erreur les plus élevés atteignant jusqu'à 66,06 % d'erreur de prédiction avec la méthode KNN à l'échelle du bâtiment. Ceci s'explique par la forte proportion de données atypiques irréelles correspondant à des blocs entiers de bâtiments fusionnés. Si le cas d'étude de Nantes pourrait être retiré pour créer un modèle avec un pourcentage d'erreur d'environ 20 %, il est certain que les données atypiques du cas d'étude de Toulouse doivent faire partie intégrante des données d'apprentissage. Il est donc finalement décidé de conserver les sept cas d'étude pour essayer de diminuer le pourcentage d'erreur de prédiction à l'échelle du bâtiment à moins de 10 % avec la méthode de classification RandomForest.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

E. Phase 3 : optimisation de la classification typologique

E.1. Objectifs

La méthode RandomForest est un algorithme constitué de plusieurs paramètres pouvant influencer les résultats de prédiction comme exposé dans la thèse de Brostaux (2005) :

- L'effectif d'apprentissage : plus sa taille est grande, plus faible est le pourcentage d'erreur de prédiction ;
- Le bruit de fond aléatoire : plus les données sont typiques vis-à-vis des classes, plus faible est le pourcentage d'erreurs ;
- Les variables non pertinentes : plus leur proportion par rapport aux autres variables est élevée, plus élevé est le pourcentage d'erreur ;
- Le nombre de variables présélectionnées lors de la construction des partitions au sein de chaque arbre : en fonction du jeu de données, le pourcentage d'erreur peut augmenter ou diminuer selon le nombre de variables présélectionnées ;
- Le nombre d'arbres constituant la forêt : en règle générale, plus il est grand, plus le pourcentage d'erreur est faible.

L'effectif d'apprentissage de chaque arbre est obtenu par tirage aléatoire avec remise d'autant de bâtiments que ceux des données typiques et atypiques des sept cas d'études. Il est donc déjà très important puisque l'on tire aléatoirement avec remise 18359 bâtiments, il n'est donc pas nécessaire de faire varier ce paramètre. Le bruit de fond aléatoire est en partie engendré par nos données atypiques que l'on souhaite prédire au même titre que les données typiques, on ne peut donc pas diminuer ce bruit de fond. Néanmoins, il nous est possible de déterminer si toutes les variables utilisées sont pertinentes, ainsi que de faire varier le nombre de variables présélectionnées (jusqu'à présent fixé à 2) et le nombre d'arbres de décision (jusqu'à présent fixé à 500 car recommandé par Culter *et al.*, 2007). Nous allons donc déterminer l'optimum de ces trois paramètres un à un afin d'optimiser la classification typologique par RandomForest à son maximum.

E.2. Méthode

L'optimisation s'est déroulée en trois temps : (1) identification des variables non pertinentes, (2) optimisation du nombre de variables présélectionnées, (3) optimisation du nombre d'arbres constituant la forêt.

Identification des variables non pertinentes

La fonction RandomForest sous R nous permet d'obtenir l'importance des variables selon deux algorithmes : Accuracy et Gini. Les variables de plus grande importance sont celles jouant un rôle majeur dans la classification des données. Les variables les moins pertinentes pour la classification sont dans notre cas celles dont l'importance est moindre. Pour déterminer si une variable est non pertinente : on construit un modèle, on identifie la variable ayant le moins d'importance d'après les deux algorithmes, on construit un deuxième modèle sans cette variable et on compare la qualité de prédiction des deux modèles. Si le deuxième modèle est meilleur ou de qualité équivalente au premier, alors la variable retirée est non pertinente.
Disposant dans notre cas de 84 variables, nous avons construit 83 modèles en retirant entre chaque modèle la variable ayant la plus faible importance. Ainsi le premier modèle est constitué de 84 variables, le deuxième de 83 variables,… et le dernier de 2 variables (utiliser une méthode de classification sur une seule variable n'ayant aucun sens). Pour chaque modèle, les pourcentages d'erreur moyens de classification des bâtiments et des typologies majoritaires des USR ont été obtenus par validation 70/30 comme exposé en C.2.

Optimisation du nombre de variables présélectionnées

Une fois les variables non pertinentes retirées, 9 modèles ont été construits avec un nombre de variables présélectionnées **mtry** respectifs allant de 1 à 9. La valeur maximale 9 correspond au **mtry** proposé par défaut dans la fonction RandomForest calculé selon le code R : `floor(sqrt(M))`, avec M le nombre total de variables utilisées pour la classification. Ce code signifie que l'on calcule l'entier inférieur de la racine carrée du nombre total de variables. Ne sachant pas combien de variables non pertinentes seront retirées, on a conservé M=84 pour

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

obtenir le **mtry** maximal de 9. Pour chacun des 9 modèles, les pourcentages d'erreur moyens de classification des bâtiments et des typologies majoritaires des USR ont été obtenus par validation 70/30 comme exposé en C.2. Le nombre de variables présélectionnées retenu sera celui du modèle ayant les plus faibles pourcentages d'erreur moyens.

Optimisation du nombre d'arbres

Une fois les variables non pertinentes retirées et le nombre de variables présélectionnées optimal déterminé, quinze modèles ont été construits avec un nombre d'arbres **ntree** respectifs de 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 600, 700, 800, 900 et 1000. Pour chacun des quinze modèles, les pourcentages d'erreur moyens de classification des bâtiments et des typologies majoritaires des USR ont été obtenus par validation 70/30 comme exposé en C.2. Le nombre d'arbres retenu sera celui du modèle ayant les plus faibles pourcentages d'erreur moyens.

E.3. Résultats

Identification des variables non pertinentes

Les pourcentages d'erreur moyens à l'échelle de l'USR suivant les mêmes tendances que ceux à l'échelle du bâtiment (figure 10) et n'étant que la conséquence des erreurs de prédiction à l'échelle du bâtiment, nous nous sommes focalisés sur les résultats à l'échelle du bâtiment. Le pourcentage d'erreur moyen le plus faible obtenu à l'échelle du bâtiment est de 11,512 % pour le modèle 12 (Figure 10). Ce modèle est construit avec l'ensemble des 84 variables moins les 11 variables de moindre importance d'après les algorithmes Accuracy et Gini de la méthode RandomForest ; ces variables sont d'après le tableau 12 : « u_VEG_NBPAI », « i_ROOF », « u_ROAD_NBPAI », « u_WATER_DENS », « u_bHOLES_A_MEAN », « u_WATER_A », « i_LRATIO_CVX, « u_WATER_L », « b_HOLES_A », « i_LRATIO_3M » et « i_L_3M ».

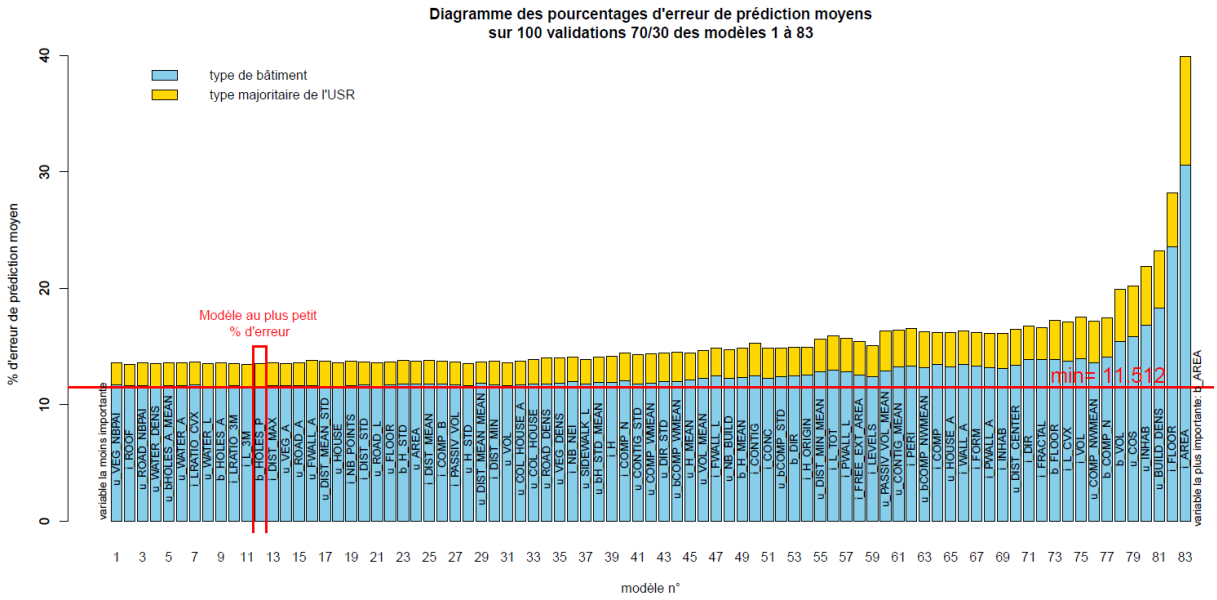


Figure 10 : Diagramme des pourcentages d'erreur de prédiction moyens sur 100 validations 70/30, à l'échelle du bâtiment et à l'échelle de l'USR.
Pour chaque modèle, la variable la moins importante est indiquée. Cette variable la moins importante n'est pas prise en compte pour l'apprentissage des modèles suivant.

FINANCÉ PAR ANR
Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Tableau 12 : Pourcentages d'erreur de prédiction moyens sur 100 validations 70/30 à l'échelle du bâtiment et à l'échelle de l'USR des modèles 1 à 83.
La variable de plus grande importance non mentionnée dans le tableau est « b_AREA *».*

modèle n°	nombre de variables prédictives	% d'erreur de prédiction de la typologie du bâtiment	% d'erreur de prédiction de la typologie majoritaire de l'USR	variable la moins importante (retirée dans les modèles suivant)
1	84	11.684	13.6	u_VEG_NBPAI
2	83	11.673	13.436	i_ROOF
3	82	11.641	13.598	u_ROAD_NBPAI
4	81	11.54	13.547	u_WATER_DENS
5	80	11.661	13.59	u_bHOLES_A_MEAN
6	79	11.657	13.591	u_WATER_A
7	78	11.682	13.655	i_LRATIO_CVX
8	77	11.554	13.569	u_WATER_L
9	76	11.516	13.626	b_HOLES_A
10	75	11.66	13.524	i_LRATIO_3M
11	74	11.539	13.472	i_L_3M
12	73	11.512	13.529	b_HOLES_P
13	72	11.654	13.602	i_DIST_MAX
14	71	11.63	13.515	u_VEG_A
15	70	11.672	13.623	u_ROAD_A
16	69	11.617	13.796	u_FWALL_A
17	68	11.59	13.726	u_DIST_MEAN_STD
18	67	11.6	13.64	u_HOUSE
19	66	11.625	13.742	i_NB_POINTS
20	65	11.681	13.657	i_DIST_STD
21	64	11.599	13.628	u_ROAD_L
22	63	11.744	13.702	u_FLOOR
23	62	11.765	13.799	b_H_STD
24	61	11.794	13.73	u_AREA
25	60	11.765	13.794	i_DIST_MEAN
26	59	11.811	13.713	i_COMP_B
27	58	11.723	13.663	i_PASSIV_VOL
28	57	11.634	13.502	u_H_STD
29	56	11.855	13.694	u_DIST_MEAN_MEAN
30	55	11.731	13.718	i_DIST_MIN
31	54	11.651	13.612	u_VOL
32	53	11.741	13.743	u_COL_HOUSE_A
33	52	11.81	13.855	u_COL_HOUSE
34	51	11.788	14.005	u_ROAD_DENS
35	50	11.822	14.01	u_VEG_DENS
36	49	11.958	14.075	i_NB_NEI
37	48	11.797	13.875	u_SIDEWALK_L
38	47	11.909	14.104	u_bH_STD_MEAN
39	46	11.94	14.143	i_H
40	45	12.07	14.443	i_COMP_N
41	44	11.806	14.282	u_CONTIG_STD

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

42	43	11.863	14.391	u_COMP_WMEAN
43	42	11.979	14.437	u_DIR_STD
44	41	12.015	14.491	u_bCOMP_WMEAN
45	40	12.133	14.468	u_H_MEAN
46	39	12.297	14.636	u_VOL_MEAN
47	38	12.503	14.879	i_FWALL_L
48	37	12.296	14.747	u_NB_BUILD
49	36	12.338	14.892	b_H_MEAN
50	35	12.481	15.26	i_CONTIG
51	34	12.24	14.85	i_CONC
52	33	12.379	14.887	u_bCOMP_STD
53	32	12.502	14.972	b_DIR
54	31	12.533	14.922	i_H_ORIGIN
55	30	12.828	15.638	u_DIST_MIN_MEAN
56	29	12.95	15.923	i_L_TOT
57	28	12.827	15.704	i_PWALL_L
58	27	12.543	15.438	i_FREE_EXT_AREA
59	26	12.426	15.082	i_LEVELS
60	25	12.918	16.334	u_PASSIV_VOL_MEAN
61	24	13.236	16.424	u_CONTIG_MEAN
62	23	13.319	16.543	i_PERI
63	22	13.161	16.29	u_bCOMP_NWMEAN
64	21	13.449	16.227	i_COMP
65	20	13.241	16.173	u_HOUSE_A
66	19	13.437	16.314	i_WALL_A
67	18	13.31	16.23	i_FORM
68	17	13.207	16.122	i_PWALL_A
69	16	13.147	16.103	i_INHAB
70	15	13.394	16.503	u_DIST_CENTER
71	14	13.883	16.764	i_DIR
72	13	13.862	16.652	i_FRACTAL
73	12	13.861	17.288	b_FLOOR
74	11	13.76	17.148	i_L_CVX
75	10	13.928	17.538	i_VOL
76	9	13.579	17.197	u_COMP_NWMEAN
77	8	14.129	17.454	b_COMP_N
78	7	15.462	19.942	b_VOL
79	6	15.839	20.228	u_COS
80	5	16.804	21.87	u_INHAB
81	4	18.29	23.25	u_BUILD_DENS
82	3	23.562	28.172	i_FLOOR
83	2	30.593	39.894	i_AREA

Pour vérifier que le modèle 12 est le modèle au plus faible pourcentage d'erreur moyen à l'échelle du bâtiment, donc que ces variables pré-citées soient bel et bien non pertinentes, et donc valider leur retrait du modèle de classification, les pourcentages d'erreur moyens très proches des modèles 8 à 21 ont été récalculés. La validation 70/30 étant basé sur des tirages aléatoires, plus on effectue de tirages (donc de modèles) plus l'estimation du pourcentage d'erreur moyen sera précise. Ainsi pour pallier le bruit de fond dû à la variabilité des pourcentages

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

d'erreur moyens calculés sur 100 modèles, on a recalculés les pourcentages d'erreur moyens des modèles 8 à 21 avec 200 modèles de validation 70/30.
On peut ainsi voir sur la figure 11 que le modèle au plus faible pourcentage est d'erreur moyen à l'échelle du bâtiment n'est pas le modèle 12, mais le modèle 11 avec 11,52 %. Les variables finalement non pertinentes et donc retirées de la classification sont donc : « u_VEG_NBPAI », « i_ROOF », « u_ROAD_NBPAI », « u_WATER_DENS », « u_bHOLES_A_MEAN », « u_WATER_A », « i_LRATIO_CVX, « u_WATER_L », « b_HOLES_A » et « i_LRATIO_3M ».

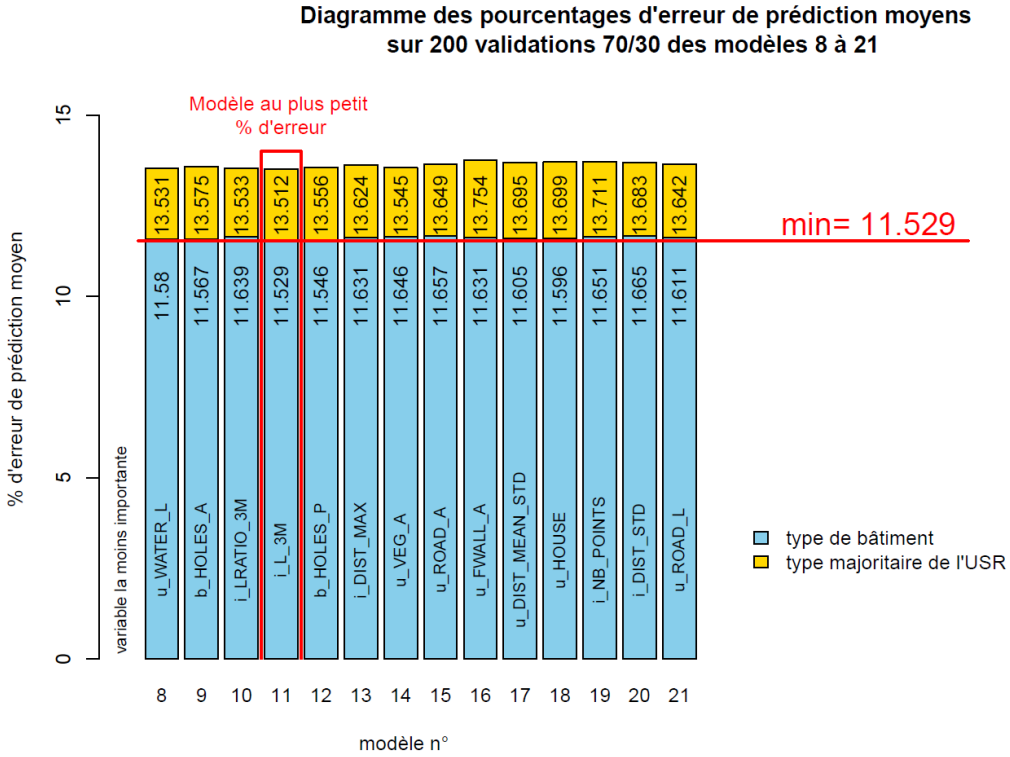


Figure 11 : Diagramme des pourcentages d'erreur de prédiction moyens sur 200 validations 70/30 des modèles 8 à 21, à l'échelle du bâtiment en bleu et à l'échelle de l'USR en jaune.
Pour chaque modèle, la variable la moins importante est indiquée. Cette variable la moins importante n'est pas pris en compte pour l'apprentissage des modèles suivant. Les valeurs de pourcentage d'erreur moyen de prédiction des bâtiments et de la typologie majoritaire de l'USR sont également indiqués.

Optimisation du nombre de variables présélectionnées

Le nombre de variables présélectionnées (mtry) à chaque nœud de chaque arbre influence les résultats de prédiction de la classification typologique. En augmentant ce paramètre, le pourcentage d'erreur de prédiction moyen à l'échelle du bâtiment diminue jusqu'à atteindre un seuil à partir de **mtry**=7 (Figure 12). Avec un modèle paramétré à **mtry**=7, on obtient alors un modèle avec 11,06 % d'erreur de prédiction moyen à l'échelle du bâtiment et 12,52 % d'erreur de prédiction moyen à l'échelle de l'USR. On décide donc de fixer la valeur du paramètre **mtry** à 7 pour notre modèle de classification typologique urbaine.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

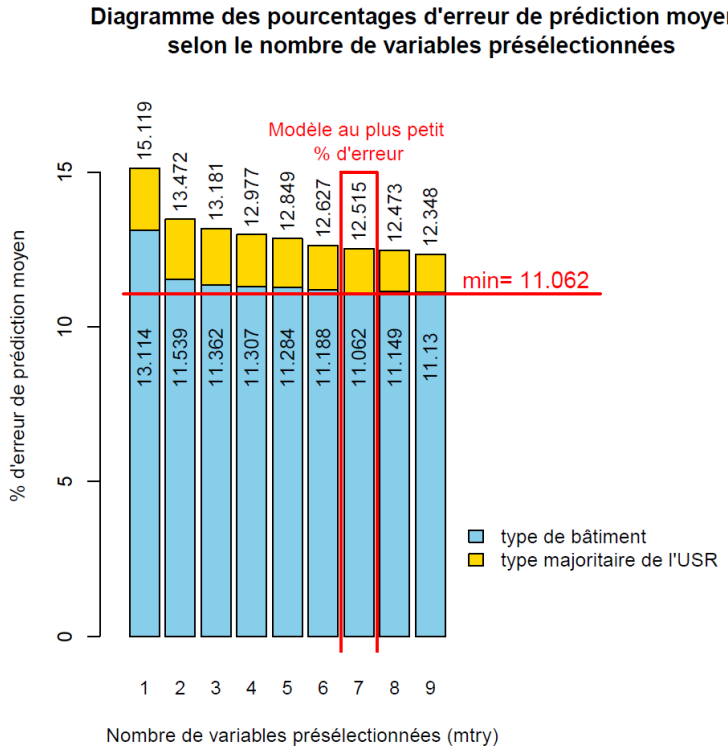


Figure 12 : Diagramme des pourcentages d'erreur de prédiction moyens sur 100 validations 70/30 selon le nombre de variables présélectionnées, à l'échelle du bâtiment en bleu et à l'échelle de l'USR en jaune. Les valeurs de pourcentage d'erreur moyen de prédiction des bâtiments et de la typologie majoritaire de l'USR sont indiqués.

Optimisation du nombre d'arbres

Le nombre d'arbres de décision constituant la forêt influence les résultats de prédiction de la classification typologique. En augmentant ce paramètre, le pourcentage d'erreur de prédiction moyen à l'échelle du bâtiment diminue jusqu'à atteindre un seuil à partir de **ntree**=500 (Figure 13). Avec un modèle paramétré à **ntree**=500, on obtient alors un modèle avec 11,06 % d'erreur de prédiction moyen à l'échelle du bâtiment et 12,52 % d'erreur de prédiction moyen à l'échelle de l'USR. Au-delà de 500 arbres, les pourcentages d'erreur de prédictions sont légèrement au dessus du seuil à cause d'un surapprentissage : trop d'arbres signifie une trop grande précision d'identification des bâtiments typiques, et donc une moins bonne prédiction des bâtiments atypiques. On décide donc de fixer la valeur du paramètre **ntree** à 500 pour notre modèle de classification typologique urbaine.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

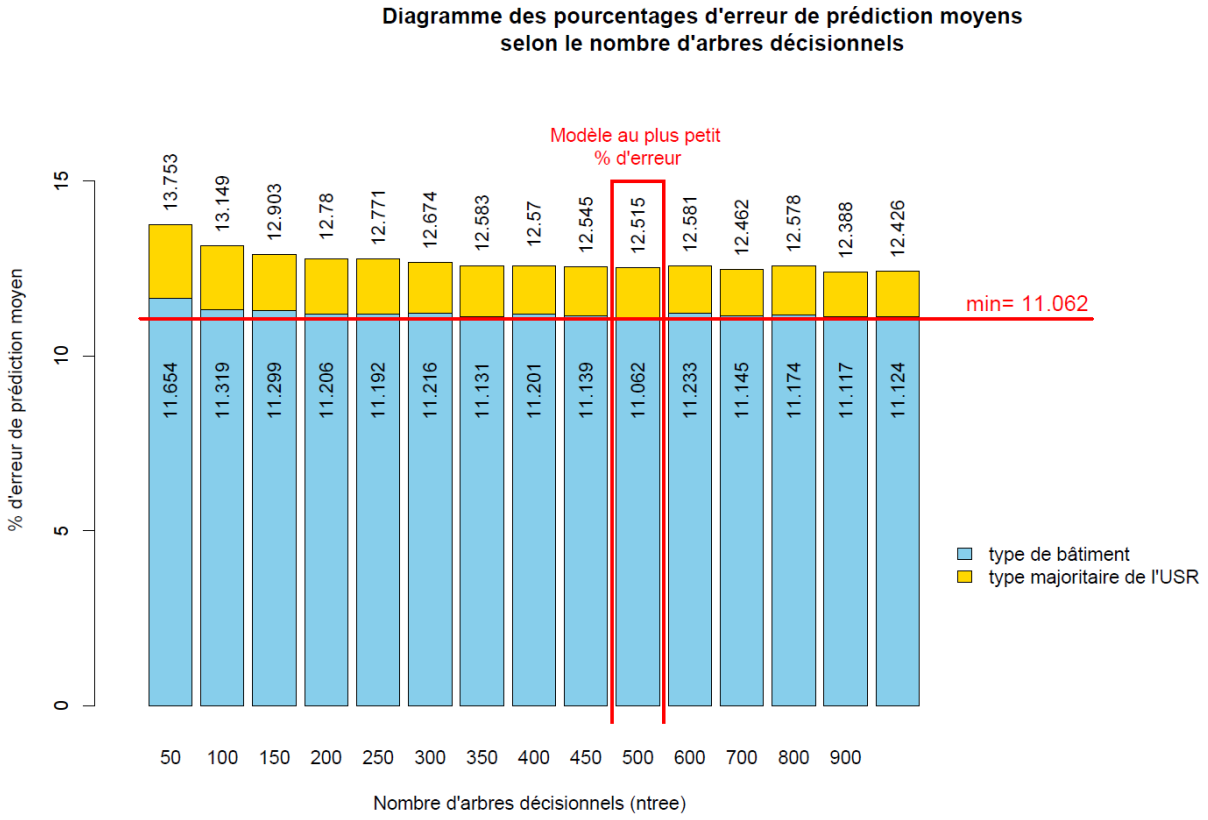


Figure 13 : Diagramme des pourcentages d'erreur de prédiction moyens sur 100 validations 70/30 selon le nombre d'arbres décisionnels, à l'échelle du bâtiment en bleu et à l'échelle de l'USR en jaune.
Les valeurs de pourcentage d'erreur moyen de prédiction des bâtiments et de la typologie majoritaire de l'USR sont indiqués.

F. Discussion

F.1. TUFA : modèle final optimisé

Après optimisation des différents paramètres, notre classification typologique urbaine basée sur la méthode randomForest utilise 18359 bâtiments (individus) et 74 indicateurs morpho-environnementaux (variables prédictives) listées dans le tableau 13. Le nombre d'individus tirés aléatoirement avec remise pour constituer le sous-échantillon d'apprentissage de chaque arbre est fixé à 18359. Le nombre de variables présélectionnées tirées aléatoirement à chaque nœud de chaque arbre est fixé à 7. Enfin, le nombre d'arbres de classification constituant la forêt est fixé à 500. Ce modèle a un pourcentage d'erreur de 11,06 % à l'échelle du bâtiment et de 12,52 % à l'échelle de l'USR d'après une moyenne sur 100 validations 70/30. On est ainsi capable de prédire la typologie urbaine d'un bâtiment en France à 88,94 % avec ce modèle baptisé TUFA pour Typologie Urbaine prédite par Forêt Aléatoire.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Tableau 13 : Liste des variables retenues pour le modèle de classification typologique urbaine TUFA.

Echelle du bâtiment	Echelle du bloc	Echelle de l'USR
i H_ORIGIN	b AREA	u VEG_A
i INHAB	b FLOOR	u ROAD_A
i H	b VOL	u ROAD_L
i LEVELS	b H_MEAN	u SIDEWALK_L
i AREA	b H_STD	u INHAB
i FLOOR	b COMP_N	u HOUSE
i VOL	b HOLES_P	u COL_HOUSE
i COMP_B	b DIR	u HOUSE_A
i COMP_N		u COL_HOUSE_A
i COMP		u FLOOR
i FORM		u COS
i CONC		u COMP_NWMEAN
i DIR		u COMP_WMEAN
i PERI		u CONTIG_MEAN
i WALL_A		u CONTIG_STD
i PWALL_L		u DIR_STD
i PWALL_A		u H_MEAN
i NB_NEI		u H_STD
i FWALL_L		u PASSIV_VOL_MEAN
i FREE_EXT_AREA		u AREA
i CONTIG		u VOL
i PASSIV_VOL		u VOL_MEAN
i FRACTAL		u NB_BUILD
i DIST_MIN		u DIST_MIN_MEAN
i DIST_MEAN		u DIST_MEAN_MEAN
i DIST_MAX		u DIST_MEAN_STD
i DIST_STD		u bH_STD_MEAN
i NB_POINTS		u bCOMP_NWMEAN
i L_TOT		u bCOMP_WMEAN
i L_CVX		u bCOMP_STD
i L_3M		u DIST_CENTER
		u BUILD_DENS
		u VEG_DENS
		u ROAD_DENS
		u FWALL_A

La méthode de classification par forêt aléatoire est encore peu utilisée dans certains domaines car n'a été développée que récemment par Breiman (2001). Mais cette méthode est de plus en plus utilisée, notamment en architecture pour identifier des typologies urbaines. En effet, la comparaison de plusieurs méthodes de classification pour prédire des typologies urbaines a déjà été réalisée par Hetch *et al.*, (2013) qui ont eux comparé les méthodes CART, KNN, BAGGING, SVM et Random Forest. Les résultats obtenus se sont alors avérés identiques aux nôtres : la méthode du Random Forest est la meilleure méthode de classification des bâtiments à partir de leurs caractéristiques morphologiques.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

F.2. Limites du modèle TUFA.

Le modèle TUFA n'est précis qu'à 88,94 %, il est donc important de noter que le modèle fait des erreurs de prédiction qui peuvent avoir des conséquences pour la suite du projet MapUCE, notamment pour estimer les consommations d'énergie et de climat urbain. Pour évaluer ces conséquences il est impératif d'identifier les erreurs commises ; c'est pourquoi nous avons extrait la matrice de confusion moyenne sur 100 validations 70/30 du modèle pour les typologies à l'échelle du bâtiment (Annexe 1) et et les typologies majoritaires à l'échelle de l'USR (Annexe 2). Pour plus de lisibilité ces matrices de confusions ont été transformées en tables de pourcentage de type observé par type prédit (tableaux 14 et 15), nous permettant ainsi de calculer pour chaque type l'erreur de déficits (erreur de déficits = pourcentage d'individus d'une classe observée affectés à d'autres classes par prédiction).

Erreurs à l'échelle du bâtiment

Tableau 14 : Table des pourcentages de type de bâtiment observé par type de bâtiment prédit.
Les pourcentages ont été calculés à partir d'une matrice de confusion moyenne sur 100 validations 70/30.

		Observations									
		ba	bgh	icif	icio	id	local	pcif	pcio	pd	psc
Prédictions	ba	84,24	0,2	1,41	4,59	8,04	1,33	0,36	0,33	0,79	0,34
	bgh	0,05	95,81	0,16	0,55	0,55	0,01	0	0	0	0
	icif	1,65	0,85	88,36	2,38	0,89	0,46	3,96	0,02	0	0,02
	icio	2,64	1,59	5,62	86,44	5,42	0,09	0,88	1,59	0,14	0,51
	id	2,66	1,48	0,8	2,59	79,92	0,06	0,55	1,18	1,56	1,27
	local	1,28	0,07	0,18	0,28	0,07	91,44	0,57	0,26	1,28	0,46
	pcif	0,82	0	3,15	0,38	0	1,69	89,64	0,78	0,09	1,14
	pcio	1,42	0	0,2	2,33	0,64	1,1	3,3	93,85	0,32	4,25
	pd	4,34	0	0	0,21	3,81	2,98	0,08	0,35	92,59	3,85
	psc	0,89	0	0,12	0,26	0,67	0,83	0,65	1,64	3,23	88,16

erreur de déficits		15,76	4,19	11,64	13,56	20,08	8,56	10,36	6,15	7,41	11,84

En regardant les types par ordre de précision de prédiction décroissante à l'échelle du bâtiment (Tableau 14), en excluant les valeurs < 1 résultantes de rares « mauvais » tirages aléatoires, on tire alors les hypothèses et conclusions suivantes :

- **bgh = bâtiment de grande hauteur**. C'est le type de bâtiment le mieux prédit avec une précision de 95,81 %, soit 4,19 % d'erreur de déficits. Les « bgh » mal prédits sont alors affectés aux types « icio » (1,59 %) et « id » (1,48 %), ce qui signifie que ces quelques « bgh » ont une hauteur trop faible pour être considérés comme des « bgh ». Le modèle TUFA permet dans ce cas de trancher sur des bâtiments qui pourraient être classés en « bgh », « id » ou « icio », et améliore donc l'objectivité de notre typologie. Ces quelques erreurs impliquent donc peu, voire aucune conséquence.

- **pcio = pavillon continu sur îlot ouvert**. Type prédit avec une précision de 93,85 %, soit 6,15 % d'erreur de déficits. Les quelques « pcio » mal prédits sont alors affectés aux types « psc » (1,64 %), « icio » (1,59 %) ou « id » (1,18 %). Les types « pcio » et « psc » étant très proches, leur confusion ponctuelle est donc normale et sans conséquence importante. La confusion avec les types « icio » et « id » peut en revanche susciter plus d'interrogations pour calculer les consommations énergétiques

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

(chauffage collectif ou individuel ?). Cette confusion est due aux « pcio » fusionnés qui sont plus proches d'une typologie immeuble que pavillon.

- **pd = pavillon discontinu**. Type prédit avec une précision de 92,59 %, soit 7,41 % d'erreur de déficits. Les « pd » mal prédits sont affectés aux types « psc » (3,23 %), « id » (1,56 %) et « local » (1,28 %). Les bâtiments de type « psc » fusionnés étant très ressemblant aux « pd », et certains « pd » étant mitoyens à un garage ou une autre annexe, leur confusion ponctuelle est normale et sans conséquence importante. Les « pd » prédits en « local » correspondent aux pavillons les plus petits sans annexe mitoyenne ou fusionnée. Pour ces quelques bâtiments (qui sont parmi les plus petits) la consommation énergétique risque donc d'être sous-estimée. A l'inverse les « pd » prédits en « id » sont les plus grands en surface et en volume. Leur affectation en « id » entrainera potentiellement une sur-estimation de consommation énergétique.

- **local = local annexe**. Type prédit avec une précision de 91,44 %, soit 8,56 % d'erreur de déficits. Les bâtiments de type « local » mal prédits sont affectés aux types « pd » (2,98 %), « pcif » (1,69 %), « ba » (1,33 %) et « pcio » (1,1 %). Ces « local » mal prédits correspondent principalement à des granges ou des hangars de la taille d'un pavillon. Leur affection au type « ba » est sans conséquence puisque les plus petits « ba » sont également des hangars ou des locaux de stockage. Leur affection au types « pd », « pcif » et « pcio » va en revanche entraîner une sur-estimation de leur consommation énergétique.

- **pcif = pavillon continu sur îlot fermé**. Type prédit avec une précision de 89,64 %, soit 10,36 % d'erreur de déficits. Les « pcif » mal prédits sont affectés aux types « icif » (3,96 %) et « pcio » (3,3 %). Les « pcif » les plus grands (en hauteur ou en volume) sont souvent divisés en plusieurs logements et devraient donc être afectés au type « icif ». Les « pcif » prédits en « icif », les « pcif » fusionnés exceptés, ne sont donc pas des prédictions fausses dans la réalité. Les « pcif » sont prédits en « pcio » lorsque les îlots sont de grande taille et les distances entre entre blocs de bâtiments plus importantes. Cette erreur de prédiction n'est donc pas une abérration dans la réalité et améliore l'objectivité de notre typologie.

- **icif = immeuble continu sur îlot fermé**. Type prédit avec une précision de 88,36 %, soit 11,64 % d'erreur de déficits. Les « icif » mal prédits sont affectés aux types « icio » (5,62 %), « pcif » (3,15 %) et « ba » (1,41 %). Les « icif » prédits en « icio » sont soit des « icif » fusionnés, soit des icif de grande taille (grande surface et grande hauteur). Il est vrai que ces « icif » de grande taille ont une grande surface d'enveloppe extérieure non mitoyenne facilitant leur ventilation, mais cela n'est valable que pour les étages supérieurs. Pour les étages les plus bas, l'îlot est fermé et la présence d'un îlot de chaleur très probable. Ces 5,62 % d' « icif » prédits en « icio » ne doivent donc pas être négligés pour estimer les incertitudes de prédiction du climat urbain où l'effet îlot de chaleur pourrait être sous-estimé. Pour les « icif » prédits en « pcif », les conséquences sont négligeables en terme de climat urbain mais la consommation énergétique de ces bâtiments pourra être sous-estimée. Enfin, les « icif » prédits en « ba » sont le résultat des USR de grande taille contenant plusieurs quartiers de différents types, notamment des « ba » et des « icif », qui auront alors les mêmes valeurs pour certains indicateurs. La consommation énergétique de ces bâtiments sera donc mal évaluée.

- **psc = pavillon semi-continu**. Type prédit avec une précision de 88,16 %, soit 11,84 % d'erreur de déficits. Les « psc » mal prédits sont affectés aux types « pcio » (4,25 %), « pd » (3,85 %), « id » (1,27 %) et « pcif » (1,14 %). Les « psc » consituent un type intermédiaire entre les pavillons discontinus et les pavillons continus. Il est donc normal et sans conséquence que le modèle prédit certains « psc » en « pd », « pcio » ou « pcif ». Les « psc » prédits en « id » sont des « psc » fusionnés qui deviennent morphologiquement proches des « id ». Cette erreur due aux données atypiques irréelles aura pour conséquence de fausser l'estimation de la consommation énergétique de ces bâtiments.

- **icio = immeuble continu sur îlot ouvert**. Type prédit avec une précision de 86,44 %, soit 13,56 % d'erreur de déficits. Les « icio » mal prédits sont affectés aux types « ba » (4,59 %), « id » (2,59 %), « icif » (2,38 %) et « pcio » (2,33 %). Les « icio » de faible hauteur (1 à 2 niveaux) et de grande surface

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

au sol ont un profil morphologique proche de certains « ba » entraînant leur confusion. Ces 4,59 % d' « icio » prédits en « ba » verront donc leur consommation énergétique et leur impact climatique mal évalués. Les « icio » prédits en « id » ne consituent aucune gène car ces deux types sont très ressemblant. Les « icio » prédits en « icif » correspondent à des « icio » de petite surface au sol, à la morphologie proche des « icif ». Les estimations climatiques pourront donc être impactées pour ces bâtiments. Enfin les « icio » prédits en « pcio » consituent probablement les habitats intermédiaires dont la morphologie est identique aux « pcio » fusionnés. Leurs usages étant similaires, cela n'aura pas ou pas de conséquences sur les consommations énergétiques.

- **ba = bâtiment d'activité**. Type prédit avec une précision de 84,24 %, soit 15,76 % d'erreur de déficits. Les « ba » mal prédits sont affectés aux types « pd » (4,34 %), « id » (2,66 %), « icio » (2,64 %), « icif » (1,65 %), « pcio » (1,42 %), « local » (1,28 %). Les « ba » de petite surface au sol ont en effet une morphologie similaire aux « pd » de plain pied, d'où une mauvaise affection de certains « ba » en « pd ». Parmi les « id » on trouve les églises et monuments religieux à la morphologie proche de certains « ba » (1 seul niveau et grande surface au sol) ; c'est pourquoi certains « ba » peuvent être affectés au type « id ». Pour les « ba » prédits en « icio », « icif » et « pcio », il s'agit de commerces de proximité situés au sein des îlots qui sont donc très contigus à l'inverse de la majorité des « ba ». Les usages étant différents entre ces typologies, les estimations de consommation énergétique et de climat urbain seront faussées pour ces bâtiments mal prédits. Pour les « ba » prédits en « local », qui sont les « ba » les plus possible en surface au sol, les conséquences sont négligeables car il s'agit ici d'un ajustement causé par l'objectivité de notre typologie qui réaffecte des petits bâtiments de stockage en locaux annexes qui consomment autant d'énergie.

- **id = immeuble discontinu**. Type le moins bien prédit avec une précision de 79,92 %, soit 20,08 % d'erreur de déficits. Les « id » mal prédits sont affectés aux types « ba » (8,04 %), « icio » (5,42 %) et « pd » (3,81 %). Les « id » sont les bâtiments aux formes les plus diverses, et donc les plus difficiles à définir morphologiquement. Ils comprennent les églises et cathédrales ainsi que des bâtiments du secteur tertiaire dont la morphologie ressemble aux « ba » (1 seul niveau et grande surface au sol) ; « ba » et « id » sont souvent réunis dans de grandes USR correpondantes à des zones d'activité et donc de nombreuses caractéristiques communes à l'échelle de l'USR ; c'est pourquoi autant d' « id » peuvent être affectés au type « ba ». Les usages étant très différents, ces mauvaises prédictions auront un impact sur les estimations de consommation d'énergie et de microclimat urbain. « id » et « icio » sont très ressemblant en terme d'usages, les « id » prédits en « icio » sont là encore le fruit de l'objectivité de la méthode qui réaffecte des bâtiments pour lesquels un *a priori* anthropique n'est pas assez précis. Ces mauvaises prédictions n'auront donc aucune conséquence. Enfin, les « id » prédits en « pd » sont des immeubles de petite taille proches morphologiquement des « pd ». Ces petits « id » peuvent cependant être du secteur tertiaire et donc avoir un usage autre que celui d'habitation. Là encore des biais sur les estimations de consommation d'énergie et de climat urbain sont potentiels.

Pour résumer les conséquences potentielles sur les estimations de consommation énergétique et de climat urbain à l'échelle du bâtiment dues à de mauvaises prédictions :

- 5,62 % d' « icif » prédits en « icio » => sous-estimation de l'effet d'îlot de chaleur urbain
- 2,38 % d' « icio » prédits en « icif » => sur-estimation de l'effet d'îlot de chaleur urbain
- 4,59 % d' « icio » et 8,04 % d' « id » prédits en « ba » => fausses estimations
- 12,71 % de « ba » prédits en différents types d'habitation => fausses estimations
- 3,81 % d' « id » prédits en « pd » => fausses estimations

Dans les 11,06 % d'erreur de prédiction du modèle, les prédictions ennoncées ci-dessus jugées perturbantes pour les estimations de consommation d'énergie et de climat urbain correspondent à 3,66 %, les autres 7,4 % d'erreur de prédiction sont sans conséquence importante.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Conséquences à l'échelle de l'USR

A l'échelle de l'USR, la mauvaise prédiction d'un bâtiment n'aura aucune conséquence dans les îlots homogènes puisque le type majoritaire restera le même. En outre ces îlots homogènes sont généralement constitués de bâtiments typiques sur lesquels il est rare d'observer des erreurs de prédictions. La mauvaise prédiction d'un bâtiment aura en revanche beaucoup plus d'impact dans les îlots hétérogènes, en particulier dans les îlots ou plusieurs types présentes une surface de plancher équitable. Ce phénomène s'observe notamment pour les îlots hétérogènes en « id » et «ba» : 6,33 % d'USR de type majoritaire « id » sont prédites en type majoritaire « ba » ; 7,15 % d'USR de type majoritaire « ba » sont prédites en type majoritaire « id » (tableau 15). Il s'observe également pour les îlots hétérogènes en « icif » et « icio » : 6,97 % d'USR de type majoritaire « icif » sont prédites en type majoritaire « icio » ; 2,58 % d'USR de type majoritaire « icio » sont prédites en type majoritaire « icif ».

Tableau 15 : Table des pourcentages de type de bâtiment majoritaire d'une USR observé par type de bâtiment majoritaire d'une USR prédit.
Les pourcentages ont été calculés à partir d'une matrice de confusion moyenne sur 100 validations 70/30.

		Observations									
		ba	bgh	icif	icio	id	local	pcif	pcio	pd	psc
Prédictions	ba	84,86	0,75	0,86	2,09	6,33	0	0,06	0,07	0,01	0,78
	bgh	0,75	92,43	0,86	0,91	0,67	0	0	0	0	0
	icif	0,32	2,9	90,24	2,58	1,23	0	3,37	0	0	0
	icio	4,41	1,73	6,97	89,26	10,34	0	4,2	7	1,21	2,42
	id	7,15	2,19	0,51	2,7	76,06	0	1,67	2,35	2,29	0,63
	local	0	0	0	0	0	0	0	0	0	0
	pcif	0,3	0	0,56	0	0,02	0	90,01	1,06	0	0
	pcio	0	0	0	2,14	1,82	0	0,68	87,9	2,55	0,22
	pd	1,54	0	0	0,06	1,47	0	0,01	0,22	93,23	6,56
	psc	0,66	0	0	0,25	2,05	0	0,01	1,41	0,71	89,4
erreur de déficits		15,14	7,57	9,76	10,74	23,94	na	9,99	12,1	6,77	10,6

La mauvaise prédiction d'un bâtiment peut également entraîné une mauvaise prédiction du type majoritaire d'une USR lorsque celle-ci n'est composée que d'un seul bâtiment. Il est courant qu'un immeuble discontinu soit seul dans une USR, donc s'il est mal prédit, le type majoritaire l'est également. Cela peut s'observer dans le tableau 15, où 10,34 % des « id » sont prédits en « icio » à cause à la fois des USR ne contenant qu'un seul « id », mais également des USR hétérogènes en « id » et « icio ». Cette observation est aussi valable pour un bloc de bâtiments fusionnés seul dans une USR. C'est le cas pour les « pcio » fusionnés en bloc qui sont prédits en « icio » ; ainsi 7 % des USR de type majoritaire « pcio » sont prédites en type majoritaire « icio ». La fusion de bâtiments est également source d'erreur pour les USR de type majoritaire « psc » dont 6,56 % sont prédites en type majoritaire « pd ».

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

F.3. Conclusion.

Le modèle TUFA, Typologie Urbaine prédite par Forêt Aléatoire, est capable de prédire la typologie des bâtiments de France avec 88,94 % de précision. Les erreurs de prédictions commises sont pour plus de la moitié sans conséquence pour l'estimation de la consommation d'énergie et du climat urbain dans la suite du projet MApUCE.
Plusieurs perspectives d'amélioration du modèle sont à explorer vis-à-vis des données et des indicateurs. Outre le fait d'ajouter des données relatives à d'autres cas d'études français pour compléter l'apprentissage du modèle et sa précision sur tout le territoire français, il serait optimal de s'affranchir des données atypiques iréelles en les corrigeant. Cette correction pourrait se faire en mettant en place une identification et une transformation automatiques de ces cas atypiques de la BD TOPO pour obtenir des représentations polygonales représentatives de la réalité ; mais elle pourrait également être faite à la source de la BD TOPO par l'IGN. Au niveau des indicateurs morphologiques, il serait intéressant d'en ajouter de nouveaux non plus seulement liés directement à la géométrie, mais liés à des paramètres topologiques (comme la dimension fractale d'un bâtiment dont le calcul peut être nettement amélioré) ou encore liés à des paramètres inter-USR tenant compte de l'influence contextuelle des USR les unes par rapport aux autres. Il est en effet important de ne pas oublier que la typologie d'une USR résulte à plus grande échelle de l'historique du quartier ainsi que du tissus urbain de la ville à laquelle l'USR appartient.
Au-delà de ces améliorations sur nos données et indicateurs, d'autres sources de données existent et pourraient alimenter le modèle TUFA. Un bon exemple est OpenStreetMap (www.openstreetmap.org) en accès en licence libre et dont les données étendues à un territoire pourraient être fusionnées à nos données nationales de la BD TOPO pour nous ouvrir de nouvelles possibilités de calculs d'indicateurs morphologiques.

G. Bibliographie

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., Li, S., 2013. FNN: Fast Nearest Neighbor Search Algorithms and Applications. R package version 1.1. http://CRAN.R-project.org/package=FNN.

Bonhomme, M., 2013. Mémoire de thèse « Contribution à la génération de bases de données multi-scalaires et évolutives pour une approche pluridisciplinaire de l'énergétique urbaine ».

Breiman, L., 2001. Random Forests. Machine Learning 45(1) : 5-32.

Brostaux, Y., 2005. Mémoire de thèse « Etude du classement par forêts aléatoires d'échantillons perturbés à forte structure d'intéraction ».

Cutler, D. R., Edwards, J. T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., Lawler, J. J., 2007. Random Forests for Classification in Ecology. Ecology 88(11) : 2783–2792.

Dray, S., & Dufour, A.B., 2007. The ade4 package: implementing the duality diagram for ecologists. Journal of Statistical Software 22(4): 1-20.

Hecht, R., Herold, H., Meinel, G., Buchroithner, M., 2013. Automatic derivation of urban structure types from topographic maps by means of image analysis and machine learning. In: Buchroithner, M. et al. (Eds.): 26th International Cartographic Conference.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

Liaw, A., & Wiener, M., 2002. Classification and Regression by randomForest. R News 2(3), 18-22.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2014. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4. http://CRAN.R-project.org/package=e1071.

R Core Team, 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/ (accessed 2015-12-14).

Ripley, B., 2014. tree: Classification and regression trees. R package version 1.0-35. http://CRAN.R-project.org/package=tree.

Therneau, T., Atkinson B., Ripley, B., 2015. rpart: Recursive Partitioning and Regression Trees. R package version 4.1-9. http://CRAN.R-project.org/package=rpart.

Tornay, N., 2015. Analyse architecturale des bâtiment typiques en France. Rapport tâche 1.2 projet ANR MapUCE.

Tornay, N., Bonhomme, M., Faraut S., 2015. GENIUS, a methodology to integer building scale data into urban microclimate and energy consumption modelling. 9th International Conference on Urban Climate, Toulouse France, 20-24 july 2015.

Venables, W. N., & Ripley, B. D., 2002. Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004

H. Annexes

Annexe 1 :
Matrice de confusion moyenne sur 100 validations 70/30 des typologies de bâtiment.

		Observations										
		ba	bgh	icif	icio	id	local	pcif	pcio	pd	psc	total
Prédictions	ba	546,7	0,2	7,86	24,32	34,43	8,14	1,89	2,53	5,8	2,06	633,93
	bgh	0,3	96,77	0,91	2,89	2,35	0,09	0	0	0	0	103,31
	icif	10,74	0,86	492,17	12,62	3,81	2,82	20,5	0,14	0	0,15	543,81
	icio	17,16	1,61	31,29	458,14	23,19	0,58	4,57	12,19	1,04	3,13	552,9
	id	17,29	1,49	4,46	13,75	342,04	0,35	2,83	9,05	11,5	7,8	410,56
	local	8,33	0,07	0,99	1,46	0,31	558,71	2,94	1,96	9,39	2,81	586,97
	pcif	5,32	0	17,57	2	0,01	10,33	464,35	5,96	0,67	6,99	513,2
	pcio	9,22	0	1,09	12,33	2,72	6,72	17,1	718,86	2,36	26,04	796,44
	pd	28,15	0	0	1,1	16,29	18,18	0,43	2,71	680,52	23,57	770,95
	psc	5,79	0	0,66	1,39	2,85	5,08	3,39	12,6	23,72	540,45	595,93
	total	649	101	557	530	428	611	518	766	735	613	5508

Annexe 2 :
Matrice de confusion moyenne sur 100 validations 70/30 des typologies majoritaires de bâtiment des USR.

		Observations										
		ba	bgh	icif	icio	id	local	pcif	pcio	pd	psc	total
Prédictions	ba	116,26	0,5	1,34	5,57	10,26	0	0,06	0,1	0,01	0,63	134,73
	bgh	1,03	61,93	1,34	2,42	1,08	0	0	0	0	0	67,8
	icif	0,44	1,94	140,77	6,87	2	0	3,44	0	0	0	155,46
	icio	6,04	1,16	10,88	237,44	16,75	0	4,28	10,71	1,05	1,96	290,27
	id	9,8	1,47	0,79	7,19	123,22	0	1,7	3,6	1,99	0,51	150,27
	local	0	0	0	0	0	0	0	0	0	0	0
	pcif	0,41	0	0,88	0,01	0,04	0	91,81	1,62	0	0	94,77
	pcio	0	0	0	5,69	2,95	0	0,69	134,49	2,22	0,18	146,22
	pd	2,11	0	0	0,15	2,38	0	0,01	0,33	81,11	5,31	91,4
	psc	0,91	0	0	0,66	3,32	0	0,01	2,15	0,62	72,41	80,08
	total	137	67	156	266	162	0	102	153	87	81	1211

FINANCÉ PAR ANR
Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-13-VBDU-0004