Cerema

Direction territoriale Nord-Picardie

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

PRÉFÈTE
DU
PAS-DE-CALAIS

Direction Départementale des
Territoires et de la Mer
du Pas de Calais.

Approche de la rétention foncière dans le Pas-de-Calais
Phase 2 : modélisation des comportements

Octobre 2016

Organisme commanditaire :

Direction Territoriale des Territoires et de la Mer du Pas-de-Calais (DDTM62)

–

Personne(s) référente(s) :

Gauthier Turco – DDTM 62

Bureau d'études : Cerema DTer NP :

Pilote et rédacteur

Martin Bocquet, chargé d'études (Cerema DTer Nord-Picardie, RDT/ stratégies foncières - expertise géomatique – SFEG) - Tél. : 03 20 49 62 71 - Courriel : martin.bocquet@cerema.fr

Relecteurs

Jérôme Douché, responsable du groupe (SFEG)

Antoine Herman, Chargé d'études (SFEG)

Directeur d'études

Jérôme Douché, responsable du groupe (SFEG)

Informations contractuelles :

Numéro d'affaire (SIGMA) : C15NR0118

Nature du rapport

- ☐ Intermédiaire
- ☐ Définitif

Historique des versions du document :

Version	Date	Commentaire
1	**10 sept 2016**	Première version
2	**19 sept 2016**	Relecture par J. Douche – Antoine Herman
3	**14 oct 2016**	Relecture par G. Turco - DDTM62

Maître d'ouvrage

Direction Départementale des Territoires et de la Mer du Pas-de-Calais (DDTM 62)

Références affaire / devis

Affaires n° C15NR0118

Bureau d'études : Cerema DTer NP

Visas techniques

Le chargé d'études pilote	Le responsable de groupe
Martin Bocquet	Jérôme Douché

Sommaire

Sommaire

Introduction

Cadrage de l'étude

Phase 1

La première partie de l'étude a permis de faire le point sur les leviers et déterminants de la rétention foncière. De même, cette partie analyse les pratiques de rétention foncière dans le Pas-de-Calais.

Objectifs de cette partie

A partir des données existantes, et de la méthode d'évaluation définie dans les parties précédentes, il est envisagé de modéliser le comportement de rétention foncière.

En d'autres termes, il s'agit de répondre à la question suivante :

est-il possible de prédire, en fonction des caractéristiques d'une parcelle, si elle sera conservée ou vendue dans les 5 prochaines années ?

Méthode utilisée

Cette partie de l'étude utilisera des notions d'apprentissage automatique (machine learning) pour répondre à la question. Ces méthodes d'intelligence artificielle, aujourd'hui en pleine évolution, permettent en théorie de déterminer des comportements à partir de données d'entrée.

Il s'agit donc de sélectionner et d'ajuster un modèle, tout en lui permettant d'apprendre à l'aide de données préparées pour lui.

Données et Fichiers fonciers

Ces modèles nécessitent de grandes quantités de données, fiabilisées et structurées pour arriver à déterminer des comportements récurrents. A ce titre, l'utilisation des Fichiers fonciers permet de disposer de données suffisamment fiables et structurées pour bâtir des modèles de qualité.

Une étude exploratoire

Les travaux réalisés dans cette partie ont été faits de manière exploratoire. Cette présente partie a été l'occasion de déterminer les possibilités de l'outil.

A ce titre, les résultats obtenus, plutôt moyens, ne reflètent pas forcément les possibilités des outils, plutôt prometteurs. En parallèle, sur d'autres études, le Cerema a obtenu avec les mêmes méthodes des résultats beaucoup plus conformes à la réalité.

Ce champ de recherche est amené à prendre de l'ampleur dans les prochaines années.

Méthode générale

Principe de la modélisation

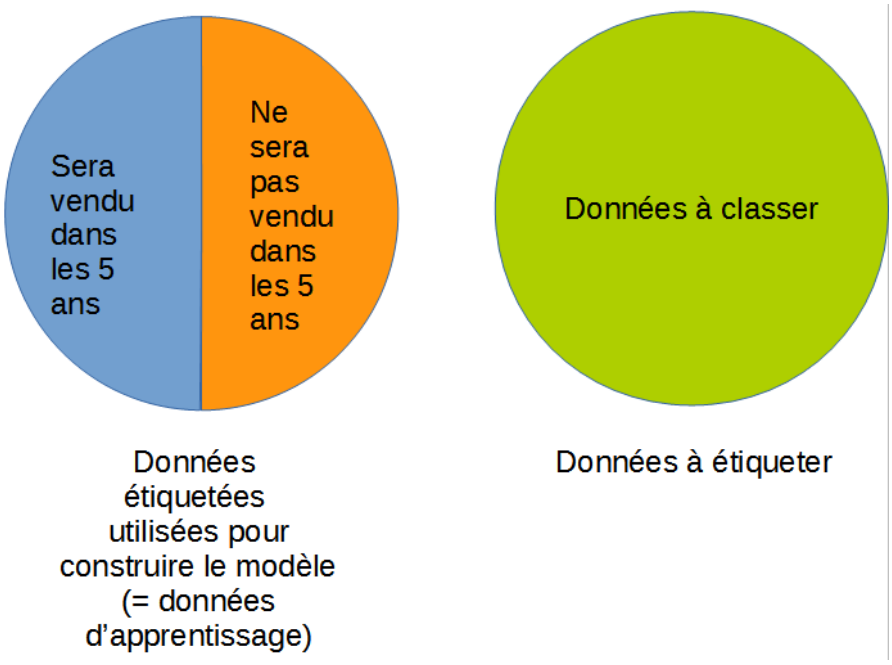
La modélisation se base sur des algorithmes d'apprentissage automatique (machine learning) dits « supervisés ».

Données de base

Chaque donnée (ici chaque parcelle) dispose de caractéristiques (taille, localisation, année de dernière mutation, etc.).

Il existe deux types de données :

- un ensemble dont on connaît le résultat. Cet ensemble de données déjà étiquetées (c'est-à-dire dont on connaît déjà la réponse) est appelé données d'apprentissage. Elles sont donc destinées à « entraîner le modèle ». Il est doté des données présentes dans chaque objet ainsi que d'une réponse (l'indicateur à modéliser, ici la question de savoir si la parcelle sera vendue ou non dans les 5 prochaines années). A partir de ces données, il est possible de créer un modèle qui permettra de connaître la destination de parcelles pour lesquelles la réponse n'est pas connue.
- Les données à étiqueter: ce sont les données pour lesquelles on cherche à prédire le résultat.

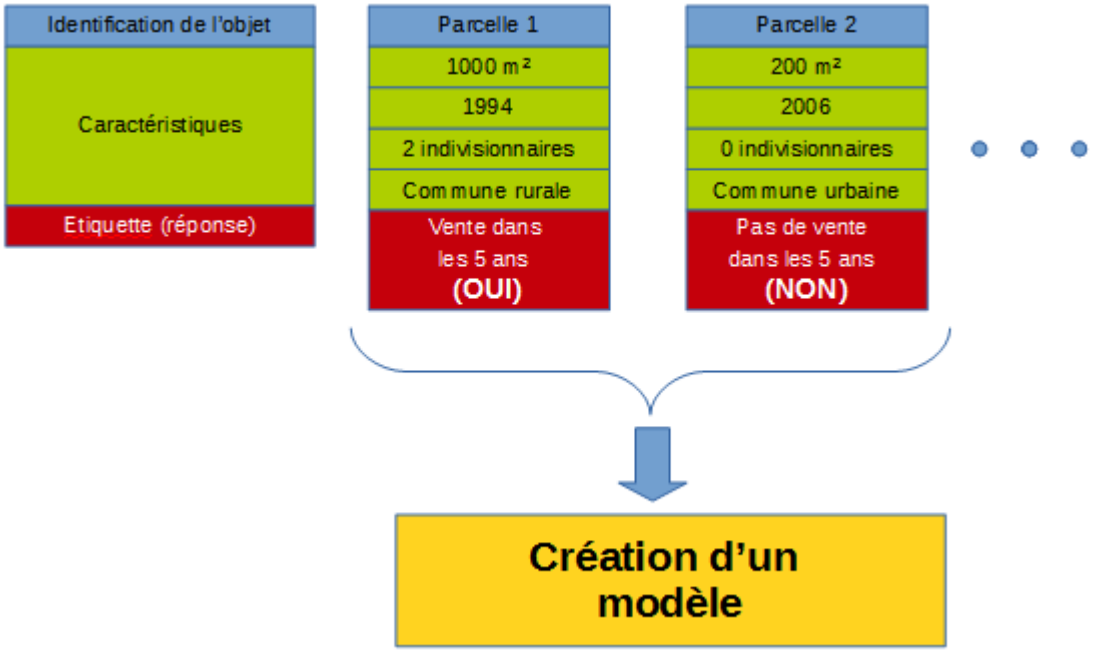


Partition des données

Modélisation et prédiction

Modélisation à partir des données étiquetées

On fournit donc les caractéristiques de chaque parcelle au modèle. L'algorithme généralisera ensuite ces caractéristiques. Par exemple, si les parcelles de plus de 1000 m² dans une commune urbaine sont toujours vendues dans les 5 ans, il est probable que toutes les parcelles présentant ces caractéristiques soient elles aussi vendues.

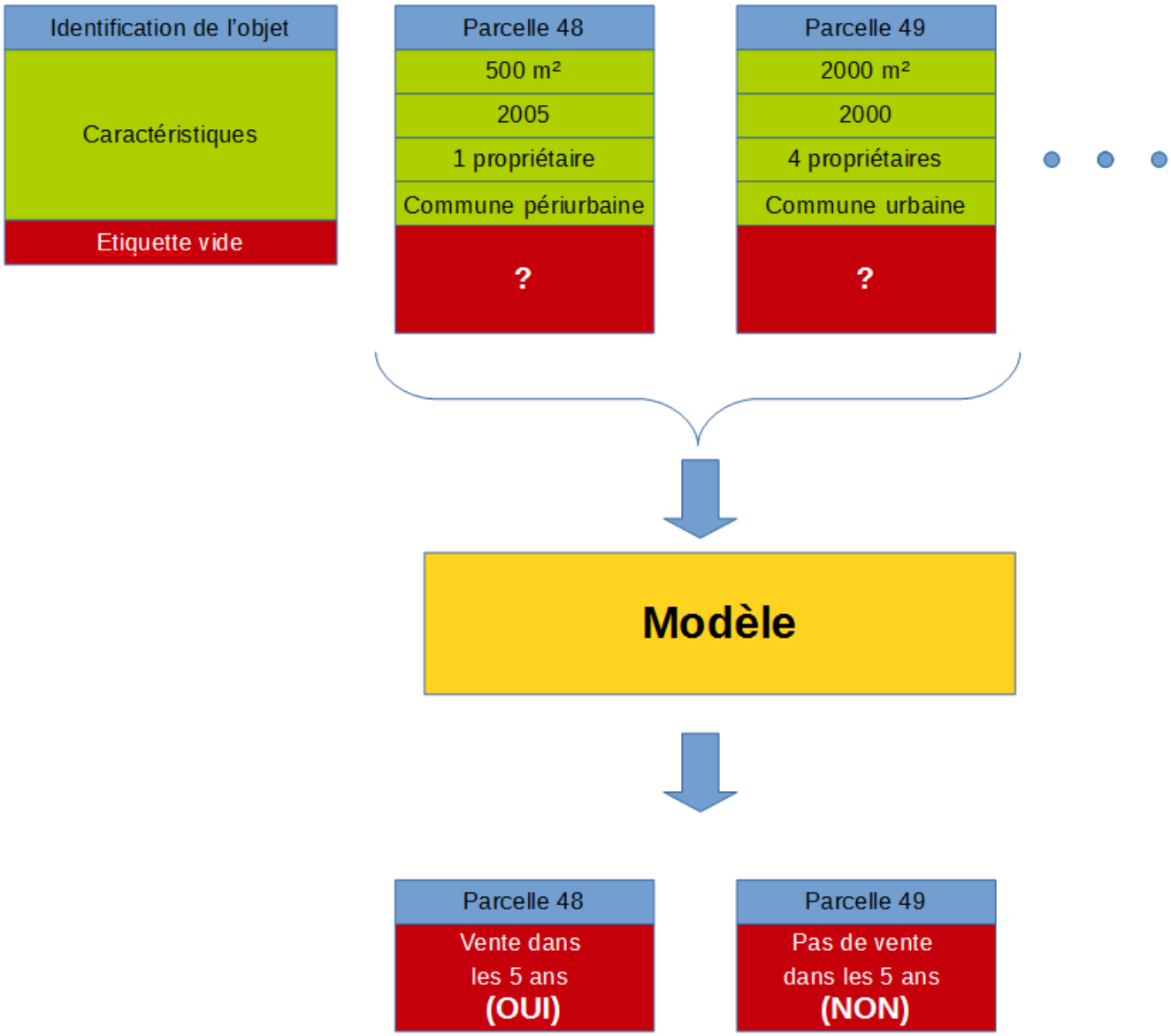


Exemple d'utilisation des données étiquetées pour créer le modèle

Prédiction du modèle sur les données non étiquetées

Une fois ce modèle déterminé, il est appliqué aux parcelles non étiquetées pour tenter de les classer.

Cette étape permet de récupérer les résultats et de définir la probabilité de vente d'une parcelle.



Exemple d'utilisation du modèle pour étiqueter les données

Validation croisée et score

Validation croisée

En réalité, seule une partie des données étiquetées est utilisée pour réaliser le modèle. L'autre partie est utilisée pour tester sa fiabilité. On réalise le modèle, et on applique ses prédictions à des données déjà étiquetées.

Les données utilisées dans la validation croisée auront donc deux champs : la réalité et la prédiction réalisée avec le modèle.

Parcelle	Classement réel	Prédiction
Parcelle 1	Vendue dans les 5 ans	Vendue dans les 5 ans
Parcelle 2	Vendue dans les 5 ans	Non vendue dans les 5 ans
Parcelle 3	Non vendue dans les 5 ans	Vendue dans les 5 ans

Cela permet de faire un premier test, à savoir le pourcentage de « bonnes » prédictions.

La matrice de confusion

Ce premier test peut cependant présenter des limites. Dans le cadre d'un échantillon déséquilibré (beaucoup plus de classements dans une catégorie que dans une autre), les scores peuvent être très hauts pour une prédiction très moyenne.

Dans le cadre de la rétention foncière, près de 75 % des parcelles ne muteront pas dans les 5 ans. Cela signifie qu'un algorithme naïf, qui prédit qu'aucune parcelle ne mutera, aura par construction 75 % de « bonnes » réponses.

Pour regarder de manière plus précise, il est nécessaire de regarder un autre indicateur : la matrice de confusion. Elle se présente sous la forme suivante :

	Prédiction: vente	**Prédiction : pas de vente**
Classement réel : vente	50 « vrai négatif »	20 « faux négatif»
Classement réel : pas de vente	30 « faux positif»	100 « vrai positif »

Exemple de matrice de confusion. 150 parcelles ont été bien classées (prédit = réalité).
30 parcelles vendues en réalité n'ont pas été prédites comme telles.
20 parcelles non vendues ont été prédites comme vendues.

En particulier, les premiers tests réalisés ont donné la matrice de confusion suivante :

	Prédiction: vente	**Prédiction : pas de vente**
Classement réel : vente	0	1880
Classement réel : pas de vente	0	5215

Cela indique que le modèle n'arrive pas à trouver les comportements liés aux ventes. En dépit de son score élevé (73,5 % de bonnes réponses), il n'apporte rien à la compréhension des phénomènes.

Algorithmes utilisés

Deux types d'algorithmes

Dans le cadre de cette étude, nous avons testé deux types d'algorithmes permettant d'effectuer des classifications binaires.

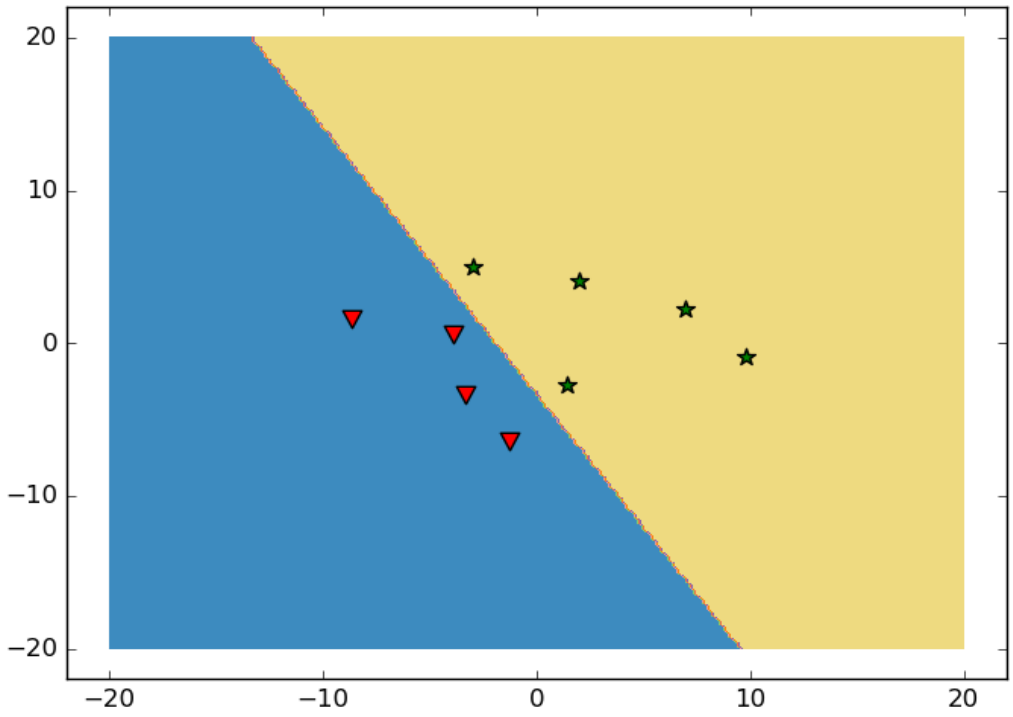
Le modèle prend donc en entrée les données, et en crée une classification, entre les parcelles vendues et non-vendues.

Ces algorithmes, programmés en Python, existent dans la librairie « scikit-learn ».

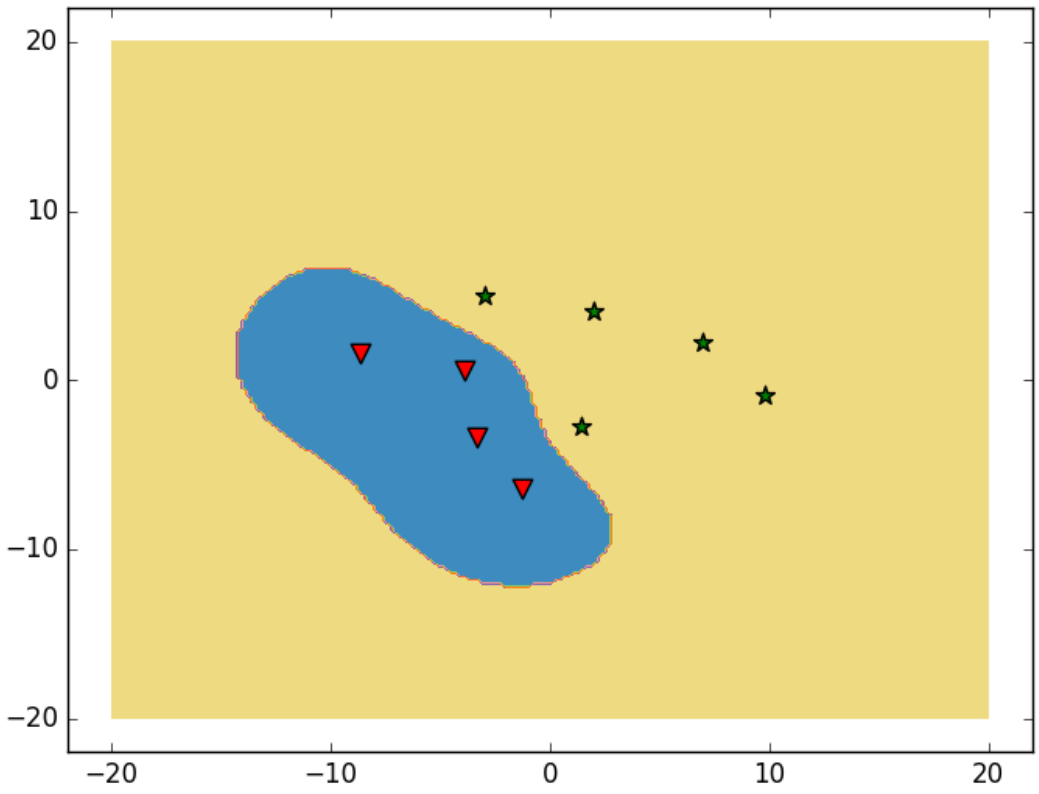
Machines à vecteur de support

Les machines à vecteur de support (aussi appelés classificateurs à vaste marge) ou « SVM » sont des algorithmes destinés à séparer géométriquement des jeux de données. Ils visent donc à construire un objet géométrique qui sépare en deux parties les données.

La séparation entre les jeux de données peut être linéaire, ou revêtir des fonctions plus complexes.



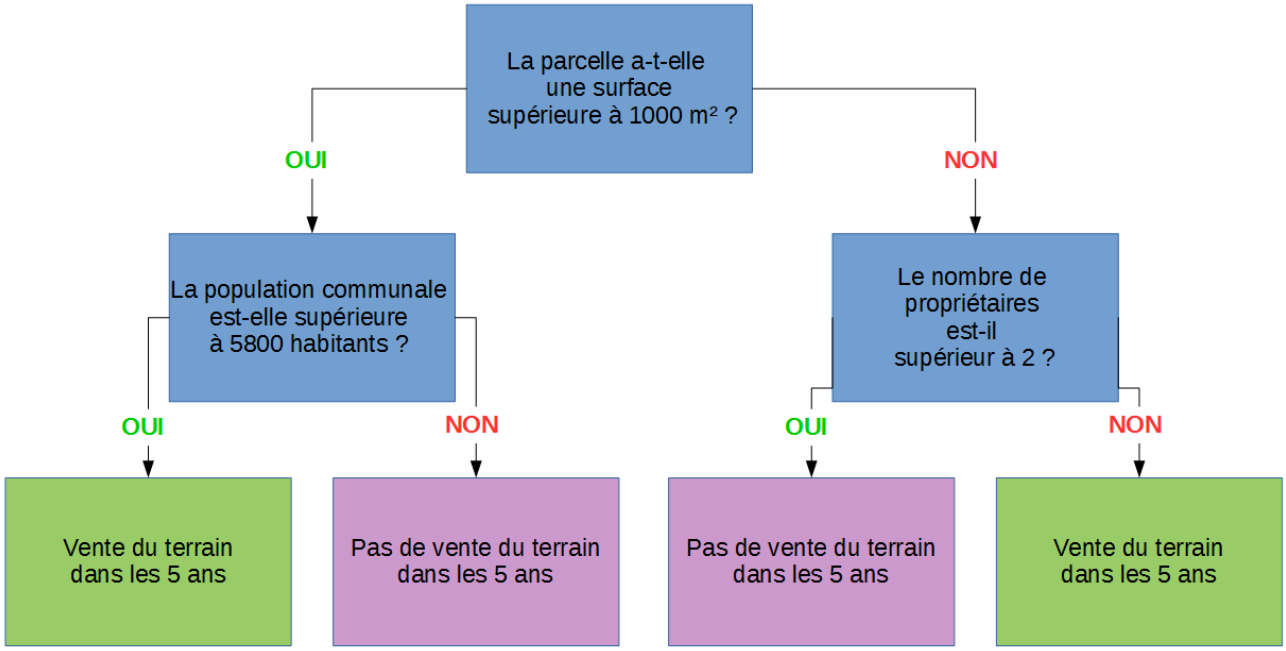
Exemple de SVM, noyau linéaire. Données aléatoires.



Exemple de SVM, noyau Gaussien. Données aléatoires.

Les forêts d'arbres décisionnels

Les forêts d'arbres décisionnels permettent de classer et déterminer les variables selon leurs paramètres. Il s'agit de créer des arbres décisionnels, capables de classer les parcelles selon leurs caractéristiques.



Exemple fictif d'arbres décisionnels

Paramètres des modèles

Les modèles développés doivent être modifiés et modulés pour permettre de meilleures performances. Par exemple, les modèles d'arbres décisionnels peuvent être entraînés en plaçant des contraintes sur leur structure : nombre de nœuds, nombre de feuilles, nombre maximal de paramètres, etc.

Une partie de l'entraînement du modèle consiste à régler ces paramètres.

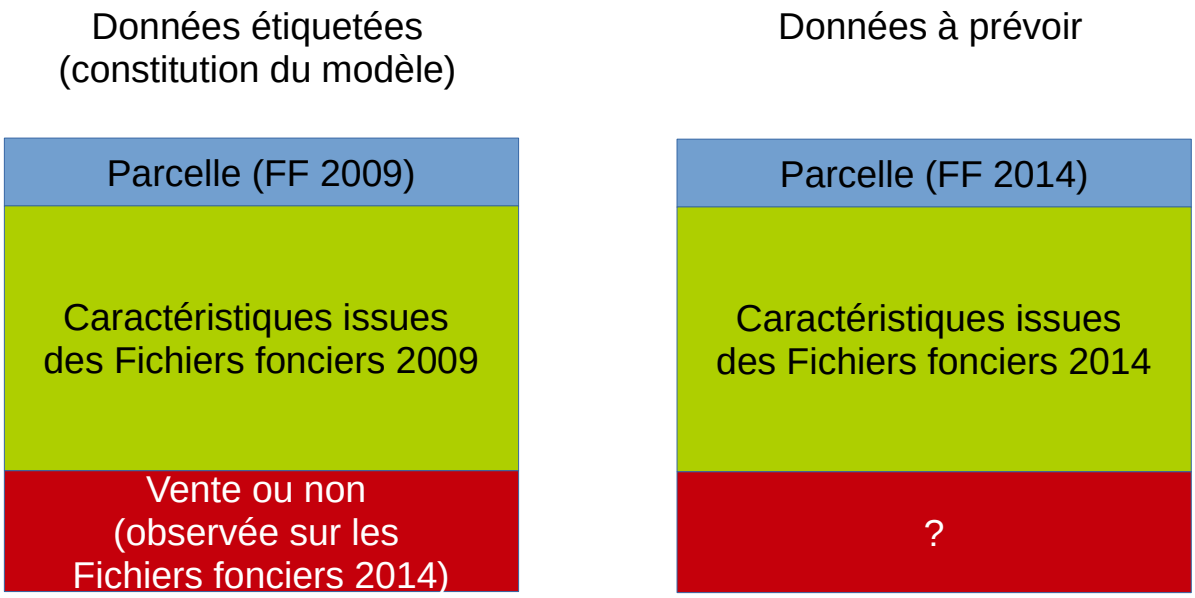
Méthode de modélisation

Objectif de la modélisation

L'objectif de la modélisation est de prévoir si une parcelle sera vendue dans les 5 prochaines années. La modélisation porte sur les parcelles libres, c'est-à-dire constructibles, selon les mêmes critères que ceux déterminés dans la partie 1.

Données étiquetées

On utilise pour les données étiquetées les Fichiers fonciers 2009. On regarde ensuite dans le millésime 2014 le devenir de ces parcelles, et si elles ont été vendues.



Provenance des données étiquetées et à prévoir

Sélection des variables

Principe

Le modèle déterminé ainsi que ses performances dépendent des caractéristiques utilisées. Il est donc nécessaire de déterminer les paramètres qui peuvent influer sur la rétention foncière.

Le choix des paramètres se fait dans un premier temps de manière très large. Des tests permettent ensuite de ne garder que les paramètres pertinents.

Paramètres choisis

Les premiers paramètres utilisés ont été déterminés dans le cadre de la phase 1.

Caractéristiques du terrain

Il a été retenu les caractéristiques suivantes. Ces critères sont issus des Fichiers fonciers (table parcelle), avec les variables suivantes :

- la surface du terrain
- la classification fiscale des sols
- l'année de dernière mutation du terrain.

Caractéristiques du propriétaire

Seuls certains critères peuvent être utilisés pour caractériser le propriétaire. Il peut s'agir de :

- son lieu de résidence (dans la commune, dans le département, etc.),
- le type de propriétaire (personne privée ou personne morale de droit privé),
- le nombre de droits de propriété,
- le nombre de droits,
- le nombre de droits d'indivision sur la parcelle.

Ces informations sont issues de la table « parcelle » et « propriétaire » des Fichiers fonciers.

Caractéristiques de la commune

L'environnement communal peut lui aussi jouer sur les probabilités de rétention foncière. Les caractéristiques communales, issues des données INSEE ont donc été ajoutées :

- la population
- la catégorie de commune dans l'aire urbaine
- la typologie Cerema de la commune (cf phase 1)
- la croissance de population entre 1999 et 2012.

Les paramètres non pris en compte

Dans le cadre de l'étude, d'autres paramètres auraient pu être testés. Ceux-ci n'ont pas pu l'être pour des raisons de temps, de disponibilité ou de fiabilité de la donnée. Il s'agit notamment :

- de l'âge du propriétaire, disponible uniquement dans les Fichiers fonciers non-anonymisés ;
- de son patrimoine foncier total, dont la reconstitution va à l'encontre de l'acte d'engagement signé pour bénéficier des Fichiers fonciers ;
- de la catégorie socio-professionnelle du propriétaire, non disponibles
- des revenus du/des propriétaires, non disponibles.

Les données liées au marché foncier (dynamisme, volume de transactions, etc.) nécessiteraient d'utiliser la base de donnée « DVF », encore en cours d'acquisition par la DGALN.

Concernant les PLU, les données les concernant n'ont pas pu être utilisées.

D'autres variables liées à l'environnement immédiat de la parcelle (taille des terrains aux alentours, accessibilité aux services, etc.) ne peuvent être utilisés facilement. Leur mobilisation sortait donc largement du cadre de l'étude.

Enfin, les données relatives aux contraintes du terrain (servitudes d'utilité publique, topographie, nuisances, pollution, etc.) n'ont pas pu être mobilisées, faute de données.

Choix des variables

Une fois les paramètres mis en place, le modèle est testé à vide (il prédit toujours la vente). Puis tous les paramètres sont testés un à un pour savoir lequel améliore le modèle.

Une fois le meilleur paramètre déterminé, on l'ajoute au modèle. Puis on recommence l'opération avec ce paramètre fixé, et on teste un à un les n-1 paramètres existants. On arrête l'opération lorsque l'ajout de nouvelles variables n'améliore plus le modèle.

Résultats

Performance globale du modèle

De manière générale, les performances globales du modèle restent modestes. L'algorithme peine à repérer correctement les parcelles qui seront vendues. Les meilleurs algorithmes obtiennent des scores à peine meilleurs que la classification naïve (score F1 de 0,72 contre 0,62 pour le modèle naïf).

Les résultats se caractérisent pas un nombre important de faux positifs (prédiction de rétention alors qu'une vente a lieu).

Cet état de fait nous permet cependant de repérer plusieurs informations intéressantes, notamment au niveau des variables.

Résultat global

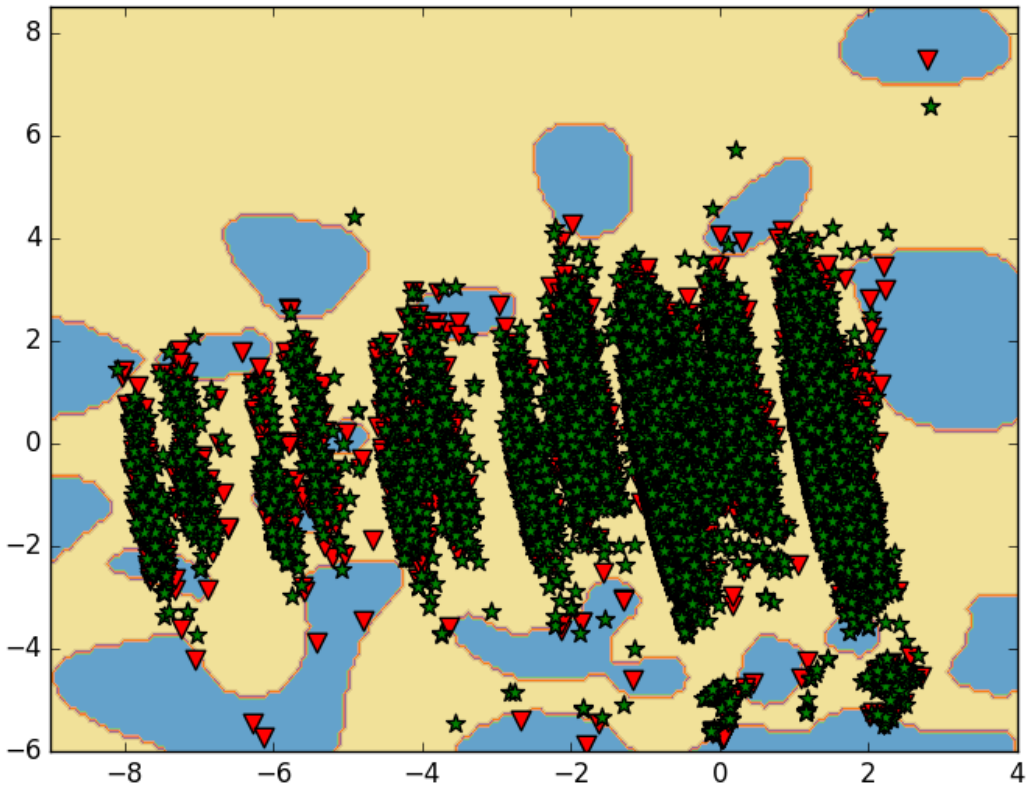
La meilleure estimation est obtenue par l'utilisation d'une forêt d'arbre décisionnel. Il est à noter que le SVM obtient des résultats similaires, mais au prix d'un temps de calcul beaucoup plus important (entre 10 secondes et 1 mn pour la forêt, entre 10 mn et 1H pour le SVM).

Dans les deux cas, la meilleure matrice de confusion ressemble à celle-ci :

	Prédiction: vente	Prédiction : pas de vente	Total du classement réel
Classement réel : vente	657	1223	1880
Classement réel : pas de vente	685	4530	5215
Total des prédictions	1342	5753	7095

Matrice de confusion du meilleur modèle obtenu. Résultat d'une forêt d'arbres décisionnels.

Lecture : sur 1880 parcelles dont la vente a eu lieu, 657 ont été prédites par le modèle. Par contre, 1223 ont été mal classées.



Graphique présentant les résultats analysés par un SVM (projection des données selon une analyse en composantes principales).

Paramètre du modèle

Ce résultat est obtenu avec une forêt d'arbres décisionnels, pour 10 arbres dans la forêt et environ 7 variables utilisées. Les données numériques ont été normalisées, et les données textuelles transformées en données binaires, pour un total de 46 variables. Les données ont été entraînées sur 28 000 parcelles, et 7095 ont été utilisées pour la validation croisée.

Interprétation

La classification est donc assez médiocre. En particulier, seules 41 % des parcelles vendues ont été prévues comme telles par le modèle. De même, 50 % des parcelles prévues comme vendues par le modèle l'ont été en réalité. Le modèle prévoit plutôt bien les décisions de rétention, puisque 87 % des parcelles en rétention ont bien été identifiées par le modèle.

Le modèle déterminé est donc une première approche, mais ne saurait être utilisé comme tel pour prédire la rétention foncière dans une commune. Au vu des faibles résultats du modèle, nous n'avons pas réalisé de cartes communales.

Explications de la rétention foncière

Les variables ont fait l'objet d'un tri, afin de connaître celles ayant le plus d'influence sur le modèle. A ce stade, et en raison des incertitudes, il n'est pas aujourd'hui possible de savoir si ces variables jouent de manière importante ou non.

Quels sont les éléments qui jouent sur la rétention foncière ?

La rétention foncière dépend en grande partie des données communales (taille de la commune, croissance, etc.), et surtout de son caractère périurbain ou non. Ainsi, les communes près d'un grand pôle, ou à cheval sur plusieurs aires urbaines, ont une rétention foncière différente.

Le nombre de propriétaires joue également de manière importante.

De même, la localisation du propriétaire est un facteur important. Le facteur clivant est la présence ou non dans la commune. En d'autres termes, un propriétaire habitant dans le même département (mais pas dans la commune du terrain) a le même comportement de rétention qu'un propriétaire d'une région lointaine.

Les caractéristiques du terrain (surface et occupation des sols) ont un rôle plus ambigu. La surface du terrain intervient à la hausse. Cependant, l'occupation telle que calculée par les Fichiers fonciers ne paraît pas intervenir de manière importante.

Conclusion et pistes d'amélioration

Pistes d'amélioration

Il pourrait être possible d'améliorer le modèle de plusieurs manières.

l'augmentation du nombre de variables

En premier lieu, on pourrait ajouter des variables explicatives. Le patrimoine, l'âge et la catégorie socio-professionnelle du propriétaire auraient pu permettre d'améliorer les performances. De plus, les variables liées au marché immobilier (volume de ventes, prix moyen, etc.) pourraient aussi être intéressantes.

Les taxes foncières, le statut du terrain (loué ou non) ainsi que son prix de location seraient d'autres variables intéressantes à ajouter.

Cependant, il faut rappeler (cf partie 1), que les décisions de ventes sont aussi liées à des décisions sociologiques non chiffrables. Par nature, le modèle ne peut donc pas s'améliorer indéfiniment.

L'augmentation du nombre d'observations

Le nombre de parcelles observées est ici d'environ 35000. Augmenter le nombre de parcelles permettrait d'améliorer le modèle. Cependant, il faut noter que :

- le nombre d'observation est déjà assez important,
- que l'augmentation du nombre d'observations passerait par l'introduction de données sur d'autres régions, ce qui pourrait présenter d'autres biais.

Il peut donc s'agir d'une piste à creuser, mais qui s'avérerait a priori moins payante.

L'amélioration du pré-traitement

L'observation de la rétention foncière s'est faite sur des parcelles qualifiées de « libres ». Cette observation est tout à fait valable à l'échelle de l'étude. Cependant, à une échelle plus fine (celle de la modélisation), il pourrait être nécessaire de raffiner la définition de parcelles « libres », notamment en incluant :

- le zonage du PLU,
- la présence de servitudes, notamment liées aux risques,
- le caractère inondable du terrain.

Connaissance et prévention des risques – Développement des infrastructures – Énergie et climat – Gestion du patrimoine d'infrastructures – Impacts sur la santé – Mobilités et transports – Territoires durables et ressources naturelles – Ville et bâtiments durables

Centre d'études et d'expertise sur les risques, l'environnement, la mobilité et l'aménagement
Direction territoriale Nord-Picardie : 44 ter rue Jean Bart - CS 20 275 - 59019 Lille Cedex
Tél : +33 (0)3 20 49 60 00 – fax : +33 (0)3 20 53 15 25
Siège social : Cité des mobilités - 25, avenue François Mitterrand - CS 92 803 - F69674 Bron Cedex - Tél : +33 (0)4 72 14 30 30
Établissement public - Siret 130018310 00016 - TVA Intracommunautaire : FR 94 130018310 www.cerema.fr