Construction d'un modèle et d'un logiciel d'explication et de prédiction des coûts immobiliers sous forme de réseau bayésien

Convention de coopération scientifique
N° SU 04 000 349 (A 04-30)
Du 09/12/2004
Tit. : U.Paris6, LIP6. Resp. : J.Y.Jaffray

FICHE-RESUME du RAPPORT FINAL

L'objectif de l'étude était la construction d'un modèle d'explication et de prédiction des coûts immobiliers sous forme de réseau probabiliste (RB) , ce modèle devant être mis à disposition sous forme de logiciel auprès de l'Agence de Développement et d'Urbanisme de l'Aire urbaine de Nancy (ADUAN).

Origine de l'étude

En 2002, l'équipe du LIP6 avait réalisé une étude préliminaire pour la DGUHC destiné à mettre en évidence les possibilités nouvelles offertes par l'utilisation des SIG pour l'observation foncière et immobilière. Les SIG permettent en effet de mettre en relation des informations variées et de dépasser les approches classiques qui ne comportaient que l'analyse des relations internes aux fichiers des transactions. Le problème est alors de disposer de méthodes d'analyse permettant de mettre en relation de nombreuses variables structurées en sous-modèles. Le LIP6 avait alors proposé d'utiliser à ces fins un outil statistique récent, les Réseaux Bayésiens (RB). Les premiers essais s'étant montré prometteurs, la DGUHC a conclu avec le LIP6 la présente convention avec pour objectif la construction d'un modèle d'explication et de prédiction des coûts immobiliers sous forme de réseau probabiliste (RB).

Le cadre du modèle : l'approche hédonique

Le modèle se situe dans le cadre de l'approche hédonique, qui a pour hypothèse que l'utilité d'un bien pour le consommateur est une fonction des diverses caractéristiques de ce bien. Un modèle hédonique cherchera donc à expliquer le prix d'un logement par sa surface, la taille des pièces, sa distance au centre-ville, etc. Cependant, la quasi-proportionnalité entre prix et surface fait que l'effet de cette variable risque de masquer celui des autres : il est donc plus intéressant de prendre comme variable expliquée le prix au m2.

Un modèle alternatif : les réseaux bayésiens

Tout d'abord nous avons choisi comme modèle les réseaux probabilistes, ou réseaux bayésiens (RB), dont la classe contient d'ailleurs, comme cas très particulier, le modèle économétrique. Les traits spécifiques de ce modèle sont les suivants : tout d'abord, il accepte sans problèmes les variables catégorielles (symboliques ; exemple : neuf/récent/ancien) ; ensuite, les liaisons entre les variables ne sont pas déterministes mais probabilistes, ce qui permet d'exprimer des faits d'observation tels que, par exemple, « toutes choses égales d'ailleurs, le prix au m2 décroît en moyenne avec la distance au centre, mais il existe des logements à la fois plus éloignés et plus chers que d'autres » ; de plus, les liaisons dues à des influences directes sont identifiées et représentées de façon à pouvoir en déduire immédiatement les liaisons indirectes;

par exemple, on pourrait dire que « les logements neufs sont en moyenne plus chers que les logements anciens, mais seulement parce qu'ils sont généralement en meilleur état ». Plus généralement la corrélation entre deux variables peut varier et même s'inverser selon les valeurs des autres variables.

Un RB est constitué : (i) d'un graphe orienté, dont les sommets sont les variables du modèle et les arcs indiquent des influences directes entre les variables à chaque extrémité et (ii) de tableaux numériques indiquant pour chacune des variables les probabilités qu'elle prenne telle ou telle valeur conditionnellement à chaque ensemble de valeurs possibles des variables parentes. La construction d'un RB et l'estimation de ses probabilités se font en général en combinant une analyse statistique des données avec du savoir d'expert.

Les renseignements que l'on peut obtenir d'un réseau bayésien sont sans commune mesure avec ce que fournirait une équation économétrique, qui permet seulement de prédire la valeur de la variable expliquée – ici, le prix – pour toute combinaison de valeurs des variables explicatives – ici, les caractéristiques des logements. Parce que l'on peut leur adjoindre un mécanisme de détection des variables dites d'actions, consistant à identifier les relations causales entre variables, les réseaux bayésiens permettent de répondre à toutes sortes de requêtes aussi bien en pronostic – « quel serait l'impact sur les prix de l'implantation de HLM dans tel ou tel quartier ? », par exemple – qu'en diagnostic, inférence inverse du pronostic – par exemple, « la différence des prix entre deux quartiers est-elle due à des facteurs sociaux ? ». Sous cet angle encore, les réseaux bayésiens présentent donc un avantage notable sur le modèle économétrique

Le système d'information

Le SIG foncier et immobilier de Nancy

Le SIG a été constitué à partir du cadastre du Grand Nancy, converti au format MapInfo par le CETE de Nantes. Sur ce référentiel, différentes couches d'information ont été constitués:

- la couche des transactions représente les parcelles ayant fait l'objet des transactions à analyser ;
- des couches issues de la base SIRENE: ensemble des établissement, extraction pour différentes catégories (enseignement, animation, etc..) ;
- des couches issues du RGP: indicateurs socio-économiques (taux d'activité, taux de chômage, composition des ménages, ...) par îlot (IRIS);
- diverses couches spécifiques créées par l'ADUAN (Francis Hess et Sylvie Chevalier): stations de tramway, de TER, parkings, périmètres ZRUZUS, espaces verts, coupures,etc..).
- le fichier des locaux du cadastre a permis, par un lien sur la référence parcellaire, l'étage et le nombre de pièces, de retrouver la description du local faisant l'objet de la transaction.

Sur l'ensemble des transactions de la période 1992-2040, plus de 30000 transactions ont pu ainsi être complètement décrites par un total de 47 variables et ont pu être traitées par la méthode des RB.

Mise en forme des données

Les études effectuées sur les prix fonciers de Nancy nécessitent une mise en forme de données très hétérogènes, celles-ci provenant de bases de données différentes stockées sous des formats différents (Access, MapInfo, etc). Afin que celles-ci soient lisibles dans le logiciel

aGRUM, il a fallu harmoniser ces données puis supprimer les transactions ayant des données manquantes ; et enfin, il faut discrétiser les variables de la base ainsi obtenue car certaines d'entre elles (comme le prix) ont trop de valeurs différentes pour être utilisables par un réseau bayésien.
Ces différents prétraitements, initialement effectués en utilisant des fichiers « texte » étaient difficiles à maintenir ; aussi a-t-on remplacé les fichiers CSV par une base PostgreSQL et les scripts shell par des scripts python. La base PostgreSQL est actuellement opérationnelle, et des scripts PHP ont été développés afin d'exporter les données dont nos logiciels aGRUM et lemon ont besoin.

La construction d'indicateurs

Motivation pour l'introduction d'indicateurs

Les RB sont d'autant plus intéressants qu'ils font apparaître des chaînes d'influence ou de causalité entre variables plus longues. Or, dans le cas d'un logement, il semble bien que l'influence d'une variable isolée sur l'appréciation globale de la qualité du logement (et donc sur son prix) soit très dépendante des valeurs d'autres variables.
Ceci suggère que les variables ayant un rôle d'intermédiaire d'influence n'agiront souvent que conjointement ; dans ce cas, il devient nécessaire de construire des indicateurs – variables synthétiques- mesurant l'influence globale de variables dont les effets ne peuvent être appréciés que conjointement ; en effet, ces indicateurs favorisent une structuration en profondeur du RB.

Identification et construction d'un indicateur

Lorsque deux logements qui valent le même prix, qui ont certaines caractéristiques communes mais diffèrent par les valeurs prises sur d'autres variables, on peut penser que ce dernier groupe de variables a une influence conjointe égale sur le prix, ce qu'on traduit en leur associant une même valeur d'un *indicateur* , fonction de ces seules variables.

Les indicateurs du modèle

Dans cette étude nous construisons trois indicateurs :
- le premier, IND_IRIS, opère la synthèse des variables relatives aux environnements socio-économiques des îlots : TXACT (taux d'activité), TXCHOM (taux de chômage), TXCADR (taux de cadres), TXOUVRIER (taux d'ouvriers), BACPLUS2ETPLUS (taux de diplômés à bac+2 au moins), ENCOURSETUDE (taux d'étudiants), TX1PERS (taux de ménages d'une personne), TX5PERSETPLUS (taux de ménages d'au moins cinq personnes) et NBPERSPARMENAGE (nombre moyen de personnes par ménage) ;
- le deuxième, IND_QL, exprime la qualité du logement en se fondant sur les valeurs des variables MAISON (maison ou non), NEUF (neuf / ancien), SURFHAB_NBPIECEPRINC (surface moyenne des pièces), CUISINE (existence / taille de la cuisine), SDB_DOUCHE (existence / nombre de salles d'eau) et ENTRETIEN (qualité de l'entretien de l'immeuble) ;
- enfin, le troisième, IND_QD, traduit une localisation plus ou moins favorable du logement, se fondant sur les valeurs des variables, renseignées grâce au SIG, DISTRUZUS, DISTLIGNTRAM et DISTECOLEPRIV indiquant respectivement les distances du logement à une ZRUZRUS, à un arrêt du tramway et à une école privée.

Développements logiciels

Nous avons développé une application fondée sur le C++. Pour ce nouveau logiciel, nous avons divisé la programmation en trois parties bien distinctes :

1. Développement d'une librairie, sans interface graphique, de manipulations de graphes (et en particulier de réseaux bayésiens) et d'algorithmes de calculs dans ces graphes. Cette librairie s'appelle aGrUM, acronyme de « a Graphical Universal Model », et compte à ce jour 57500 lignes de C++.
2. Développement d'une interface graphique en GTKmm utilisant les objets d'aGrUM. Cette librairie s'appelle LEMON (Library for Easily Modelling and Operating on Networks). Elle compte 24000 lignes de C++.
3. Développement d'un logiciel fondé sur les librairies aGrUM et LEMON capable d'échanger des informations avec des bases de données PostgreSQL. Cette application est appelée PULP (Probabilistic Updating and Learning Program).

Le réseau bayésien

Nous appuyant sur les logiciels précédents et à l'aide de l'algorithme LMN de construction de RB que nous avons mis au point, nous avons pu construire un RB ayant pour variables une liste sélectionnée sur avis d'experts complétés par des considérations statistiques et les trois indicateurs construits précédemment.

Le RB fait apparaître l'influence directe sur les prix des trois indicateurs, mais aussi de la variable ANCIMM (âge du logement au moment de la transaction), qui contient l'information *neuf / ancien,* et de la variable MAISON, qui distingue *maison / appartement.*

L'intérêt du logiciel est de pouvoir explorer les effets d'actions sur les variables. Par exemple, si l'on ne regarde que les maisons, on constate que dans le nouveau RB la variable ASCENSEUR perd toute importance ; si en revanche on ne regarde que les appartements, l'influence de MAISETA (étage) apparaît comme importante.

Conclusions de l'étude

Cette étude a permis de confirmer que :

i) Les systèmes d'information géographiques (SIG) permettent d'enrichir utilement les bases de données immobilières ;

ii) Le concept d'indicateur peut être rendu opérationnel par la création d'une méthodologie statistique appropriée ;

ii) Le modèle des réseaux bayésiens (RB) permet de mettre en évidence des influences complexes subtiles qui échappent aux modèles classiques.

Enfin et principalement, l'étude a abouti à la réalisation d'un logiciel d'explication des prix immobiliers riche et souple, offrant de nombreuses opportunités nouvelles d'analyse.