Chapitre 2 : Théorie des indices hédoniques

2.1 Le pourquoi des approches hédoniques

Les classes d'une nomenclature de biens et services, même restreintes, comportent des éléments de caractéristiques très disparates. Les logements (les biens considérés ici) se distinguent ainsi par leur localisation, leur superficie, leur nombre de pièces, le fait d'être occupés ou pas, leur ancienneté... tandis que les locations de logements (les services) se différencient en plus par les durées de bail, les possibilités de renouvellement, les clauses de sorties anticipées.

Cette hétérogénéité se retrouve sur les marchés correspondants, entraînant des taux de rotation ou d'occupation des logements variés, et des prix très différents. Elle entraîne diverses difficultés pour l'analyse des prix.

La principale vient du fait que le prix d'un logement ne peut être observé qu'au moment des transactions, assez rares (observabilité partielle) ; de façon analogue un loyer n'est observable que s'il y a utilisation du service. En dehors de ces situations, ces valeurs, prix et loyers, n'ont d'ailleurs pas d'existence au sens économique.

Une façon classique de contourner cette difficulté est de supposer des *valeurs implicites*, appelées également dans notre contexte « estimation du prix » du logement et « valeur locative ». Ces valeurs implicites ne sont connues que lorsqu'elles coïncident avec des prix de transaction ou des loyers et ne pourront donc être reconstituées qu'à partir de modèles décrivant la composition des prix ou des loyers, et leur évolution.

Les approches hédoniques reposent sur de tels modèles et expliquent comment les utiliser pour construire les valeurs non observées et définir des ensembles cohérents d'indices de prix.

2.1.1 Modèle hédonique avec stratification a priori

Nous allons présenter l'approche hédonique à partir d'une formulation simplifiée. Nous supposons que les divers biens peuvent être regroupés en strates, définies a priori et telles que les évolutions de prix soient approximativement parallèles à l'intérieur d'une même strate. Elles peuvent en revanche différer sensiblement d'une strate à l'autre.

Dans notre application, les strates seront des zones géographiques élémentaires où se trouvent les logements. Nous désignons ces strates par s, $s=1,\ldots,S$. Étant donné un logement i de la strate s, de caractéristiques $z_{i,s,t}$ (superficie, nombre de pièces, etc.), nous estimons, grâce à un modèle de régression, sa valeur implicite $p^*_{i,s,t}$ à la date t.

A l'intérieur de la strate s, les valeurs implicites sont supposées telles que :

$$p^*_{i,s,t} \approx c(s, z_{i,s,t})\, p^*_{s,t} \qquad (2.1)$$

- où $p^*_{s,t}$ est une valeur implicite de référence pour la strate s à la date t,
- $c(s,z_{i,s,t})$ un coefficient correctif tenant compte des caractéristiques du bien et dont la valeur peut dépendre de la strate,
- $\approx$ signifie « peu différent de ».

Le correctif et la *valeur de référence* sont définis à un scalaire multiplicatif près (ce qui est un problème d'identification). Il est alors possible de fixer la valeur de référence comme correspondant à un bien de qualité pré-spécifiée (logement de référence), z_0 :

$$p^*_{0,s,t} \approx p^*_{s,t} \Leftrightarrow c(s, z_0) = 1$$

disons le deux-pièces, ancien, se trouvant au rez-de chaussée, etc.

$\dfrac{p^*_{i,s,t}}{c(s, z_{i,s,t})}$ sera nommé « prix équivalent bien de référence » du logement (i,s) à la date t.

L'approximation (2.1) peut-être rendue plus proche d'un modèle économétrique en introduisant de façon appropriée des termes d'erreur et en spécifiant une forme paramétrée pour le correctif. Une telle spécification est par exemple :

$$\log p^*_{i,s,t} = \sum_{k=1}^{K} \beta_{k,s} X_k\left(z_{i,s,t}\right) + \log p^*_{s,t} + \varepsilon^*_{i,s,t} \qquad (2.2)$$

où les termes d'erreur $\varepsilon^*_{i,s,t}$ sont supposés indépendants, de moyenne nulle et de variance $\eta^2_{s,t}$ dépendant éventuellement de la strate et de la date. Les $X_k\left(z_{i,s,t}\right)$ sont des variables explicatives en nombre *K*, fonctions des caractéristiques du logement ou de croisements de telles ou telles caractéristiques.

La formulation (2.2) présente l'avantage de recourir à un modèle linéaire dans les paramètres $\hat{c}_0(s,z) = \exp\left[\sum_{k=1}^{K} \hat{\beta}_{k,s} X_k(z)\right]$, qui vont permettre de définir le correctif, et $\log p^*_{s,t}$, qui va fournir la valeur de référence de la strate.

On notera que le modèle comporte des effets croisés de la strate et des autres caractéristiques du logement, les coefficients $\beta_{k,s}$ étant spécifiques de la strate.

2.1.2 Des démarches alternatives problématiques

Peut-on reconstituer des évolutions de prix sans utiliser un modèle hédonique ? Deux démarches alternatives sont possibles : l'une, la méthode des ventes répétées, est difficile à mettre en œuvre rigoureusement, et l'autre, basée sur l'observation des prix moyens, conduit à des résultats biaisés.

Même si un bien donné *i* est rarement échangé à deux dates successives données *t*, *t+1* (on parle alors de données répétées), on peut cependant espérer mesurer des évolutions de prix en comparant deux biens analogues. Si les variables explicatives $X_k(z)$ sont par exemple qualitatives, il peut exister plusieurs biens de la strate ayant approximativement les mêmes niveaux de prix et pas seulement des évolutions parallèles. On utilisera ces prix pour calculer une évolution. Cependant, dès que les effets de qualité sont suffisamment importants et nombreux, le marché peut ne pas être assez liquide pour que ceci soit réalisable. Par ailleurs on néglige une grande partie de l'information contenue dans les données de transaction en ne retenant que les prix correspondant aux ventes répétées.

Une seconde démarche souvent proposée consiste à comparer le prix moyen des biens de la strate *s* échangés en *t+1*, au prix moyen des biens de cette même strate échangés en *t*, avec l'espoir d'approcher ainsi l'évolution de la valeur de référence. Cette approche est biaisée, car la structure par qualité des biens échangés n'est pas stable dans le temps. Pour illustrer cette difficulté, considérons un cas, où un seul bien est échangé à chaque date dans la strate, le bien étant de qualité z_t à la date *t*. Le rapport des prix observés serait approximativement :

$$\frac{c(s,z_{t+1})}{c(s,z_t)} \frac{p^*_{s,t+1}}{p^*_{s,t}}$$

et diffère de l'évolution de la valeur de référence à cause de la modification du terme correctif.

2.2 Utilisation du modèle hédonique

L'utilisation de modèles permet de pallier l'inobservabilité partielle et de corriger les effets de qualité. La démarche comprend plusieurs étapes selon le schéma ci-dessous.

- Étape 1 : estimation des coefficients correctifs grâce à un ensemble prédéfini de données de transactions, appelé dans la suite *parc d'estimation*. Choix d'un ensemble de logements, appelé *parc de référence*, dont on suivra les prix.
- Étape 2 : à chaque date *t*, utilisation des données de transactions effectives et des correctifs estimés à l'étape 1 pour reconstituer les valeurs des logements du parc de référence.

- Étape 3 : utilisation des valeurs de référence et des correctifs pour construire une batterie d'indices de prix et un système expert de valorisation.

2.2.1 Estimation des coefficients correctifs

On considère un parc d'estimation constitué d'un ensemble de transactions durant une période prédéfinie $t=1,...,T_0$ (dite *période d'estimation*). Les transactions fournissent des couples (prix $p_{j,s,t}$, qualité $z_{j,s,t}$), $j=1,...,J_{s,t}$, $s=1,...,S$, $t=1,...,T_0$, où $J_{s,t}$ désigne le nombre de transactions dans la strate s à la période t.

On estime les paramètres $\beta_{k,s}, k=1,...,K$ dans chaque strate, par moindres carrés ordinaires, grâce à l'équation (2.2). On en déduit une estimation du terme correctif de la strate :

$$\hat{c}_0(s,z) = \exp\left[\sum_{k=1}^{K} \hat{\beta}_{k,s} X_k(z)\right]$$

ainsi qu'une fourchette, éventuellement, pour chacun de ces termes, en considérant la précision des coefficients estimés :

$$\left[\hat{c}_1(s,z), \hat{c}_2(s,z)\right]$$

Ces correctifs seront conservés inchangés pendant une période future, dont la longueur doit être précisée.

2.2.2 Estimation de la valeur de référence de la date t

Considérons maintenant une date t, postérieure à la période d'estimation et des données de transaction : $p_{j,s,t}$, $z_{j,s,t}$, pour chacun des biens j de la strate s à la date t. Nous avons :

$$\begin{aligned} \log p_{j,s,t} &= \log c\left(s, z_{j,s,t}\right) + \log p^*_{s,t} + \varepsilon_{j,s,t} \\ &\cong \log \hat{c}_0\left(s, z_{j,s,t}\right) + \log p^*_{s,t} + \varepsilon_{j,s,t} \end{aligned}$$

après remplacement des correctifs par leurs approximations obtenues à la première étape.

On obtient $J_{s,t}$ estimations du prix du bien de référence de la strate s à la date t, $p^*_{s,t}$, et on en déduit l'approximation par moindres carrés ordinaires de $p^*_{s,t}$:

$$\log \hat{p}^*_{s,t} = \frac{1}{J_{s,t}} \sum_{j=1}^{J_{s,t}} \left[\log p_{j,s,t} - \log \hat{c}_0\left(s, z_{j,s,t}\right)\right]$$

$$\Leftrightarrow \hat{p}^*_{s,t} = \prod_{j=1}^{J_{s,t}} \left[\frac{p_{j,s,t}}{\hat{c}_0\left(s, z_{j,s,t}\right)}\right]^{\frac{1}{J_{s,t}}}$$

$p^*_{s,t}$ est donc estimé par la moyenne géométrique des prix équivalents bien de référence. La prise en compte des écarts-types permet également de proposer une fourchette pour la valeur de référence. Plus précisément, notons :

$\hat{\eta}^2_{s,t}$ la variance empirique des valeurs $\log\left(p_{j,s,t} / \hat{c}_0\left(s, z_{j,s,t}\right)\right), j=1,...,J_{s,t}$.

Cette variance empirique approche celle $\eta^2_{s,t}$ du terme d'erreur. La valeur de référence admet un logarithme compris entre $\log \hat{p}^*_{s,t} - 2\hat{\eta}_{s,t}$ et $\log \hat{p}^*_{s,t} + 2\hat{\eta}_{s,t}$, et la fourchette est :

$$\left(\hat{p}^*_{1,s,t} = \exp\left(-2\hat{\eta}_{s,t}\right)\hat{p}^*_{s,t}, \hat{p}^*_{2,s,t} = \exp\left(2\hat{\eta}_{s,t}\right)\hat{p}^*_{s,t}\right)$$

2.2.3 Construction d'un système de valorisation

On peut alors estimer les valeurs implicites de tout bien de la date *t*, de qualité z, non nécessairement échangé.

L'estimation est :

$$\hat{p}^*_{s,t}\hat{c}_0(s,z)$$

la fourchette peut être prise égale à :

$$\left(\hat{p}^*_{1,s,t}\hat{c}_1(s,z),\ \hat{p}^*_{2,s,t}\hat{c}_2(s,z)\right)$$

2.2.4 Construction d'une batterie d'indices

La méthode hédonique a permis de définir des indices élémentaires par strate :

$$I_{s,t} = \hat{p}^*_{s,t}, s = 1,...,S,\ t = 1,...,T$$

qui peuvent servir de base à la construction d'indices synthétiques. Cette dernière est menée de façon classique, en suivant par exemple une approche de type indice de Laspeyres. Pour cela, on définit dès la période initiale un panier de biens (un parc de logements), appelé parc de référence, en précisant les qualités *z* et strates *s* concernées. *Z* désigne les qualités différentes introduites dans le panier (parc de référence) et $N_{z,s}$ le nombre de biens (logements) de caractéristiques *z* dans la strate *s*.

L'indice synthétique est défini en suivant la valeur de ce panier :

$$I_t = \sum_{s=1}^{S}\sum_{z\in Z} N_{z,s} I_{s,t}$$

Ce panier peut aussi servir à construire des indices synthétiques désagrégés de façon cohérente. On peut par exemple introduire un indice « deux pièces », en considérant le sous-parc restreint aux seuls deux pièces :

$$I_t(\textit{deux pièces}) = \sum_{s=1}^{S} \sum_{z\in Z\ (\textit{deux pièces})} N_{z,s}\ I_{s,t}$$

ou un indice pour une région donnée, calculé sur un sous-ensemble de strates :

$$I_t(\textit{région}) = \sum_{s\in \textit{région}} \sum_{z\in Z} N_{z,s} I_{s,t}$$

De façon standard, on peut prendre par convention une année de base, *t = 0*. La normalisation doit alors être effectuée indice par indice, conduisant à des indices base *100* à *t = 0* donnés par :

$$I_{t/0} = 100\ I_t / I_0$$
$$I_{t/0}(\textit{deux pièces}) = 100\ I_t(\textit{deux pièces}) / I_0(\textit{deux pièces})$$

2.3 Rendre plus robuste l'approche hédonique

L'approche, reposant sur des estimations, peut être sensible à la précision de celles-ci ou à l'instabilité des paramètres. Il est utile de faire éventuellement quelques regroupements de variables de qualité ou de strates permettant d'accroître la significativité des résultats.

2.3.1 Recherche des scores sous-jacents

Considérons les termes correctifs. Pour chaque strate, nous avons estimé un ensemble de paramètres $\hat{\beta}_{1,s},...\hat{\beta}_{K,s}$ définissant le *score* de la strate s : $\sum_{k=1}^{K}\hat{\beta}_{k,s}X_k$.

On peut examiner, si ces *S* scores ne dépendent pas d'un nombre plus faible de scores sous-jacents. Pour cela, une démarche est la suivante :

1. On définit la matrice $\hat{B}$, de taille *(K,S)*, dont les colonnes sont les vecteurs $\left(\hat{\beta}_{1,s},...,\hat{\beta}_{K,s}\right)$;
2. On détermine les valeurs propres $\hat{\lambda}_1 \geq \hat{\lambda}_2 ... \geq \hat{\lambda}_S$ et les vecteurs propres associés $\hat{\alpha}_1,...,\hat{\alpha}_S$ de la matrice $\hat{B}'\hat{B}$, où $\hat{B}'$ désigne la transposée de $\hat{B}$.
3. Le nombre de valeurs propres S_0 significativement différentes de zéro donne le nombre de scores indépendants sous-jacents et ceux-ci sont donnés par :

$$Z_l = \hat{\gamma}_l' X = \hat{\alpha}_l' \hat{B}' X, l = 1,...,S_0$$

4. La robustification consiste à contraindre le terme correctif à être de la forme :

$$c(s,z) = \exp\left[\sum_{l=1}^{S_0} \lambda_{l,s} Z_l\right]$$

Comme S_0 est inférieur à la fois à *K* et *S*, souvent assez petit, il y a beaucoup moins de paramètres à estimer dans cette forme contrainte de terme correctif.

Quitte à désagréger en partie les scores Z_l déduits de l'approche précédente, il est usuel de retenir des sous-scores ne comportant que des variables avec des interprétations de même type. Par exemple un sous-score effectuera la correction pour les caractéristiques physiques du logement, un second pour son aménagement, un troisième pour la qualité de l'environnement, un quatrième pour sa localisation... On obtient ainsi une structure hiérarchique des effets des variables, qui facilite la mise en place des systèmes experts de valorisation, leur compréhension et leur mise à jour.

2.3.2 Regroupement de strates

De façon analogue, on peut examiner les strates par l'intermédiaire de l'évolution des indices $I_{s,t}$ correspondants. Des analyses des corrélations empiriques entre ces séries temporelles peuvent permettre de repérer des strates dont les valeurs de référence évoluent de façon parallèle. Elles peuvent alors, si ceci est interprétable, être regroupées.

2.4 Surveiller la spécification

Les valeurs des paramètres peuvent se modifier dans le temps et les correctifs, calculés à la période d'estimation, se détériorer. Il est important de mettre en place des instruments de suivi de la qualité du modèle afin de repérer le moment où il devra être réestimé et avoir des idées sur les modifications à apporter. On pourra procéder à un examen approprié des résidus d'estimation.

A chaque date *t* de tels résidus sont :

$$\hat{\varepsilon}_{j,s,t} = \log p_{j,s,t} - \log \hat{c}_0\left(s, z_{j,s,t}\right) - \log p^*_{s,t}\ , j = 1,..., J_{s,t}$$

Ils doivent être agrégés pour à la fois éliminer des effets qualités et orienter la surveillance vers les paramètres $\beta_{k,s}$ susceptibles d'être affectés. Pour cela, on peut considérer diverses caractéristiques marginales, par

exemple : deux-pièces (d.p.), logement ancien (a.), ..., et pour chacune d'entre elles moyenner les résidus correspondants[18] :

$\hat{\varepsilon}_{s,t}(d.p.) =$ moyenne des $\hat{\varepsilon}_{j,s,t}$ sur les deux pièces de la strate *s* et de la date *t*,

$\hat{\varepsilon}_{s,t}(a.) =$ moyenne analogue sur les logements anciens...

Si le modèle est bien spécifié, de telles moyennes devraient varier autour de zéro. On va donc rechercher des écarts systématiques. Par exemple constater des valeurs $\hat{\varepsilon}_{s,t}(a.)$ trop souvent positives pour la strate s_0, à partir d'une certaine date t_0, peut nécessiter la réévaluation d'un paramètre relatif à une variable explicative X_{k_0} en rapport avec l'ancienneté. La mise en évidence d'une tendance dans la suite $t \to \hat{\varepsilon}_{S_0,t}(a.)$ peut signifier que l'hypothèse de proportionnalité d'évolution des prix à l'intérieur de la strate *s* n'est plus satisfaite et que cette strate doit être décomposée.

2.5 Extension du modèle de base

Le modèle hédonique utilisé jusqu'à maintenant à titre d'illustration présente l'inconvénient de définir a priori des strates homogènes pour les évolutions des prix. Ce modèle peut être généralisé de la façon suivante.

Commençons par écrire la formulation (2.2) sous une forme résumée :

$$\log\ p^*_{i,s,t} = \sum_{s_0=1}^{S}\sum_{k=1}^{K} \xi_{s_0}(i,s,t)\beta_{k,s_0} X_k(z_{i,s,t}) + \sum_{s_0=1}^{S} \log\ p^*_{s_0,t}\ \xi_{s_0}(i,s,t) + \varepsilon^*_{i,s,t}$$

où ξ_{s_0} désigne la variable indicatrice de la strate s_0, c'est-à-dire celle qui vaut 1, si l'observation se trouve dans la strate s_0, 0 sinon. Sous cette forme, le modèle s'applique à l'ensemble des données toutes strates confondues, et reste linéaire dans les divers paramètres :

$$\beta_{k,s_0}, k = 1,...,K,\ s_0 = 1,...,S,\ \log p^*_{s_0,t},\ s_0 = 1,...,S,\ t = 1,...,T$$

Ses limites apparaissent clairement dans les expressions des variables explicatives : la partie « terme constant » inclut des effets croisés strate × qualité très spécifiques, alors que les parties donnant la dynamique des valeurs de référence : $p^*_{s_0,t}$ ne font pas intervenir les effets de qualité autres que l'indicatrice de la strate.

Un modèle élargi est alors :

$$\log p^*_{i,t} = c_0(z_{i,t};\theta_0) + \sum_{l=1}^{L} c_l(z_{i,t};\theta_l) f_{l,t} + \varepsilon^*_{i,t} \qquad (2.3)$$

où $f_{1,t},...,f_{L,t}$ sont des facteurs dynamiques à déterminer, $c_0(z_{i,t};\theta_0),...,c_L(z_{i,t};\theta_L)$ des termes correctifs donnant par exemple les sensibilités aux facteurs, avec des paramètres $\theta_0,...,\theta_L$ à estimer. Il n'y a plus lieu ici de distinguer la strate *s* qui est réintégrée parmi les autres caractéristiques du logement.

La démarche hédonique peut être appliquée à partir d'un tel modèle. Par exemple, en situation courante de la date *t*, les valeurs des facteurs seront estimées par moindres carrés ordinaires sur le modèle approché :

$$\log p_{j,t} \cong \hat{c}_0(z_{j,t}) + \sum_{l=1}^{L} \hat{c}_l(z_{j,t}) f_{l,t}$$

où les $\hat{c}_l(z_{j,t})$ ont été déterminés à la période d'estimation. Nous ne discuterons pas de façon plus approfondie les procédures d'estimation correspondantes. Signalons seulement que le modèle (2.3) permet de déterminer les meilleures prévisions des facteurs en supposant tous les prix implicites observables. Ces prévisions apparaissent

[18] Si le nombre de transactions se révèle insuffisant pour calculer de telles moyennes, on peut augmenter la période sur laquelle la moyenne est calculée en considérant deux ou trois dates consécutives.

comme des combinaisons linéaires des logarithmes de prix, avec des coefficients dont la somme peut être normalisée à 1 :

$$f_{l,t}^{*} \cong \sum_{i=1}^{N} \pi_l(z_{it}) \log p_{i,t}^{*}$$

$$\Leftrightarrow \exp f_{l,t}^{*} = \prod_{i=1}^{N} \left[p_{i,t}^{*}\right]^{\prod_{l(z_{i,t})}}$$

À chaque facteur correspond ainsi implicitement un panier (parc) de composition $(\pi_l(z_{i,t}),\ i$ variant), variant dans le temps, tel que l'évolution de f_l soit proche de celle de la valeur du panier (du parc). On parle de *panier* (parc) *mimétique* du facteur (Huberman, Kandel, Stambaugh, 1987). Ainsi il existe des choix adaptés des indices désagrégés dont les évolutions vont reproduire celles des facteurs sous-jacents.