Une reformulation informationnelle de l'indice de ventes répétées - Applications et conséquences pour la mesure du prix de marché de l'immobilier

Arnaud Simon

▶ **To cite this version:**

Arnaud Simon. Une reformulation informationnelle de l'indice de ventes répétées - Applications et conséquences pour la mesure du prix de marché de l'immobilier. Mathématiques [math]. Université Paris Dauphine - Paris IX, 2006. Français. NNT : . tel-00150905

HAL Id: tel-00150905
https://theses.hal.science/tel-00150905

Submitted on 31 May 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE PARIS-DAUPHINE
D.F.R. SCIENCES DES ORGANISATIONS
DRM-CEREG

Une reformulation informationnelle de l'indice de ventes répétées

Applications et conséquences pour la mesure du prix de marché de l'immobilier

THESE

présentée et soutenue publiquement par

Arnaud SIMON

pour l'obtention du titre de

DOCTEUR EN SCIENCES DE GESTION

Arrêté du 25 avril 2002

JURY

Directeur de recherche :	Laurent BATSCH Professeur à l'Université Paris-Dauphine
Rapporteurs :	Patrice FONTAINE Professeur à l'Université de Grenoble II
	Edwin DEUTSCH Professeur à l'Université de Technologie de Vienne
Suffragants :	Carole GRESSE Professeur à l'Université Paris-Dauphine
	Patrick ROGER Professeur à l'Université Louis Pasteur, Strasbourg I
	Jacques FRIGGIT Chargé de mission au Conseil Général des Ponts et Chaussées

Novembre 2006

L'université n'entend donner aucune approbation ou improbation aux opinions émises dans les thèses : ces opinions doivent être considérées comme propres aux auteurs.

Remerciements

Mes premiers remerciements vont évidemment au Pr. Batsch qui a accepté d'assurer l'encadrement de ce travail dans ce domaine si particulier de la Finance Immobilière. Son soutien, ses encouragements et sa confiance sont pour beaucoup dans cette thèse.

Je remercie le Pr. Fontaine et le Pr. Deutsch du grand honneur qu'ils me font en acceptant de participer au jury en tant que rapporteurs. Les commentaires qu'a faits le Pr. Fontaine au SIFF 2005 de Louvain-la-Neuve m'ont bien aidé à préciser ma problématique de recherche.

Je remercie ensuite le Pr. Ginglinger, en tant que directrice du centre de recherche du CEREG, et le Pr. Gresse en tant que membre du jury. Les commentaires particulièrement avisés du Pr. Gresse sur le document de présoutenance m'ont été très précieux.

Je remercie également le Pr. Roger et J. Friggit d'avoir accepté de participer à ce jury. Les cours dispensés par le Pr. Roger à l'Université Paris-Dauphine m'ont permis de conjuguer correctement ma formation antérieure de mathématiques et les problématiques financières, en évitant les écueils sur lesquels peuvent parfois s'échouer les étudiants en provenance des Sciences « dures ». Le livre de J. Friggit, sa réflexion et son expérience dont le domaine de l'Immobilier quantitatif m'ont également été d'un grand profit.

J'adresse un remerciement particulier à Fabrice Riva pour sa disponibilité, sa compétence et aussi pour avoir accepté de participer à la présoutenance en relisant un document de travail relativement aride et ingrat. Et cela au péril de sa vie ! « J'ai cru mourir », tels furent ses mots au sujet d'une partie mal écrite, lourde et à l'époque assez ésotérique.

Je remercie également les participants aux réunions et aux conférences de l'AREUEA (American Real Estate and Urban Economics Assocaition), de l'ERES (European Real Estate Society) et de l'AREIM (Association de Recherche et d'Etudes en Immobilier) pour tous les commentaires pertinents et substantiels qu'ils ont pu faire sur ce travail. Je pense en particulier à John Clapp, Chris Redfearn, Thomas Thibodeau, Martin Hoesli, Andrea Gheno, Kelvin Wong, Michel Baroni, Mahdi Mokrane, Fabrice Barthélémy, Bernard Thion …

Je remercie de plus tous les enseignants des Masters 104 et 246 pour la qualité de leurs cours, ainsi que les personnels administratifs du CEREG, en particulier Bernadette et Christine.

Enfin je remercie Arzé, Francois, Stelios, Khaoula, Sarah et Julie pour avoir accepté de relire divers passages de cette thèse. Je n'ai même pas eu à employer des moyens déloyaux ou machiavéliques pour qu'ils acceptent de lire de la Finance Immobilière ! Plus généralement, je remercie tous les doctorants et docteurs du CEREG à qui j'espère bien pouvoir continuer de casser les pieds en leur parlant, encore et toujours, de Finance Immobilière…

SOMMAIRE GENERAL

INTRODUCTION

CHAPITRE 1 : Un nouveau formalisme pour la méthode des ventes répétées

CHAPITRE 2 : Méthodologie d'analyse de données et étude de la fiabilité de l'indice

CHAPITRE 3 : Volatilité, réversibilité et problème des deux populations

CHAPITRE 4 : Vers une théorie des indices informationnels

CONCLUSION

ANNEXES ET BIBLIOGRAPHIE

Introduction

1. La notion de prix de marché de l'immobilier

L'objet de cette thèse, à savoir le prix de marché de l'immobilier, n'existe pas. L'ambition qui consiste à vouloir résumer à un seul chiffre (la valeur de l'indice) l'état d'un marché aussi hétérogène, illiquide et flou que l'immobilier est assez déraisonnable et très mal fondée théoriquement. Mais si du point de vue conceptuel les difficultés sont réelles, il n'en reste pas moins que ce chiffre unique est absolument incontournable ; il n'est pas possible, tant pour les praticiens de la gestion que pour les théoriciens, de s'arrêter à ce constat. C'est en gardant à l'esprit cette tension entre la rigueur scientifique et la nécessité de l'action que ce travail a été réalisé. Nous discuterons dans la suite de ce paragraphe des problèmes théoriques de la mesure du prix, au niveau d'un bien isolé puis pour le marché dans son ensemble.

Au niveau désagrégé des biens immobiliers, les difficultés de la notion de prix sont nombreuses et parler au singulier de ce concept est déjà discutable. Prenons l'exemple d'une maison située près d'une école réputée. Pour un couple avec enfants cette caractéristique est un élément de valorisation compte tenu de la carte scolaire en usage en France par contre, pour des retraités, il n'y aura pas de raison de surpayer cette situation géographique. Les valeurs que ces deux types d'agents économiques accorderont à cette maison seront donc différentes[1]. Les mécanismes d'établissement du (des) prix sont d'autre part relativement complexes pour un bien comme l'immobilier. En général les transactions sont intermédiées par un agent, on peut alors s'interroger sur l'impact qu'a ce tiers sur le prix de vente (Evans, Kolbe (2005)). Dans le cas des ventes aux enchères on peut également se poser la question, comme Ong, Lusht et Mak (2005), de savoir quels sont les principaux facteurs influençant le prix et le succès de l'enchère (prix proposé supérieur au prix de réservation du vendeur). Toutes ces questions relèvent d'un champ que l'on pourrait appeler

[1] On pourra consulter sur ce thème l'article de Leung, Leong, Wong (2006) qui cherche à déterminer l'ampleur de la dispersion des prix pour un même bien immobilier et les facteurs macroéconomiques influençant ce phénomène.

« microstructure immobilière » et elles mettent en évidence la complexité de la notion de prix dans ce secteur. Mais, à nouveau, si ce concept est discutable il a cependant le mérite d'exister. L'observation d'un prix de transaction reste une donnée irremplaçable.

Toujours au niveau désagrégé, il faut prendre en compte un élément supplémentaire, à savoir celui de la faible rotation des biens immobiliers qui sont en général détenus pendant plusieurs années. Implicitement on est alors souvent amené à estimer la valeur du stock (non échangé) par les prix constatés sur le flux (échangé), à une date donnée. Et si le flux n'est pas représentatif du stock, ce qui est tout à fait envisageable dans le cas des biens non fongibles, la détermination de la juste valeur devient un exercice périlleux. On glisse ainsi insensiblement du concept de prix au concept de valeur, rendant encore plus délicat l'exercice de la mesure monétaire requis par la finance (cf. Thion (1998))

Inévitablement, les difficultés que l'on rencontre en cherchant à définir sans ambiguïtés la notion de valeur ou de prix pour un bien particulier, se répercutent sur le concept agrégé de "prix de marché de l'immobilier". Sans aller jusqu'à affirmer qu'un marché résulte uniquement de conventions sociales comme Favereau, Biencourt et Eymard-Duvernay (2002), on peut toutefois s'interroger sur le niveau de solidité théorique de ce concept. En d'autres termes, quel est le degré de quantifiabilité ou de mesurabilité de cette notion curieuse que l'on appelle "LE prix de l'immobilier" et qui fait si souvent la une des hebdomadaires grand public ?

Pour illustrer plus concrètement ce problème, on peut considérer une situation où l'échantillon d'estimation est constitué de deux types de biens : des studios-T1-T2 typiquement destinés aux primo accédants et des appartements de grande taille correspondant mieux aux besoins et aux moyens des familles avec enfants. Si l'on considère que les caractéristiques de rendement et de volatilité sont identiques pour ces deux catégories de logements, il est alors légitime d'estimer un seul indice. Par contre, si l'on a des raisons de penser comme Clapp et Giaccotto (1999) que les dynamiques sont différentes, quel sera alors le sens économique de l'indice ? Aura-t-

il une signification réelle ou ne sera-t-il qu'un résultat mathématique découlant de l'application mécanique d'une procédure économétrique? Ce type de questionnement revient en fait à s'interroger sur l'existence, ou plus exactement la représentativité, de l'indice (cf. Case, Pollakowski, Watcher (1991)).

On pourrait répondre très rapidement en affirmant que deux indices détaillés valent toujours mieux qu'un seul. Mais si ce raisonnement est utilisé une fois pour distinguer les petits appartements des grands, pourquoi ne pas le réitérer en distinguant le plus finement possible les différents types de biens et obtenir ainsi des indices aussi spécialisés que "l'indice des studios, 3eme étage, sans ascenseur, orienté au nord, donnant sur cour" ? On obtiendrait alors une collection zoologique de courbes retranscrivant fidèlement l'hétérogénéité des biens immobiliers. Cette approche, appliquée trop naïvement[2], présente en fait le grand inconvénient de faire perdre de vue l'ambition du chiffre unique (LE cours de l'immobilier). D'autre part, plus un échantillon est homogène plus il est petit, et plus l'indice estimé est susceptible d'être affecté par une volatilité indésirable en provenance des transactions exceptionnelles. Le calcul d'un indice s'inscrira donc dans une tension entre une démarche dissociante, visant à assurer une homogénéité suffisante dans l'échantillon d'estimation, et un principe moniste requis par des nécessités économétriques mais surtout par la volonté d'avoir un seul chiffre sur lequel s'appuyer pour pouvoir prendre des décisions de gestion. Cette intentionnalité numérique commune à la plupart des méthodes de finance moderne (méthode des comparables pour l'évaluation d'entreprise, cotation, choix d'investissement à l'aide des critères VAN et TIR, pricing d'option par réplication…) est motivée par la flexibilité de la notion d'ordre total, caractéristique des nombres réels[3]. Mesurer, c'est ensuite pouvoir comparer puis décider du meilleur investissement. Il faut toutefois mentionner que la notion d'ordre total est à manier avec précaution, car comme l'a souligné Markowitz (1952) un choix d'investissement ne peut

[2] Par application naïve on entend calcul d'un indice médian et, en aucun cas, calcul d'un indice hédonique (on pourra se reporter au paragraphe suivant pour une présentation de ces différentes méthodologies).

[3] De deux nombres réels on pourra toujours dire lequel est le plus grand

pas se faire à la seule vue du rendement espéré, on est obligé de raisonner en termes de couples **{**Rendement espéré ; Risque de l'investissement**}**. Dans ce genre de situation la notion d'ordre total disparaît au profit d'une notion d'ordre partiel[4]. Il n'en reste pas moins que la mesure du prix de marché de l'immobilier (l'indice) est la première étape pour l'étude des décisions d'investissement.

Comme nous pouvons le constater, les marchés très imparfaits tels que l'immobilier et le marché de l'art, caractérisés principalement par leur illiquidité et leur hétérogénéité, positionnent donc, de facto, la recherche dans ce domaine près des frontières de la finance moderne. Beaucoup de choses restent à inventer dans cette Terra Incognita…

La suite de ce chapitre introductif, dont la fonction est de présenter le contexte général dans lequel s'inscrit cette thèse consacrée à l'indice de ventes répétées, est organisée de la façon suivante. Le deuxième paragraphe présente les différentes méthodologies indicielles : indice médian, indice sur valeurs d'expertise, indice hédonique, indice de ventes répétées, indice hybride. Le paragraphe trois discute des enjeux de ce sujet pour l'allocation d'actifs et pour le développement des produits dérivés sur indices immobiliers qui semblent rencontrer actuellement un certain succès. Les sections quatre et cinq mettent en rapport le thème des indices avec les prêts immobiliers (gestion actif-passif bancaire) et les problématiques économiques et sociales liées au logement (politiques urbaines). Enfin, le paragraphe six présentera la problématique de recherche de cette thèse et l'articulation des différents chapitres.

[4] L'ordre partiel s'applique typiquement à l'ensemble des nombres complexes.

2. La diversité des méthodologies indicielles en immobilier

Même si "le prix de l'immobilier" est une chimère, ce concept reste d'une très grande utilité en pratique. Il faut donc pouvoir le définir le plus explicitement possible, quitte à avoir recours à un certain degré de convention. On expose ici succinctement les bases des principales méthodologies développées pour répondre à cette question ; les deux notions centrales qui les sous-tendent sont l'illiquidité et l'hétérogénéité.

Ces différentes techniques sont très couramment employées. Elles concernent des situations souvent très diverses, tant du point de vue sectoriel que du point de vue géographique. Ainsi Chau, Wong, Yiu, Leung (2005) étudient le marché résidentiel de Hong-Kong ; Nappi-Choulet, Maleyre, Maury (2005) le marché des bureaux de la Défense ; Sunderman, Spahr, Birch, Oster (2000) le prix des ranchs aux Etats-Unis. Les séries de prix immobiliers peuvent être parfois très longues : Friggit (2001) remonte par exemple au XIX^{eme} pour le prix des logements en France, Eichholtz (1997) étudie les variations de prix le long du canal Herengracht en Hollande de 1628 à 1973, tandis que Eitrheim et Erlandsen (2004) couvre la période 1819-1989 pour quatre villes de Norvège.

Sur le plan du vocabulaire et de manière un peu transversale, on distinguera deux familles d'indices : les indices de prix (médian, hédonique, indice sur valeurs d'expertise…) et les indices de rendement (indice de ventes répétées). Dans le premier cas, la valeur de l'indice à une date t ne sera qu'une moyenne, plus ou moins sophistiquée, des prix enregistrés à cette même date. Le niveau de l'indice à t ne dépendra donc pas du passé ou du futur des prix immobiliers. Dans le deuxième cas la « moyenne » portera sur les rendements et, comme on le verra ultérieurement, la structure temporelle du RSI[5] sera très différente. La valeur à l'instant t sera en fait une fonction des prix d'achat (avant t) et des prix de revente (après t).

[5] RSI : « repeat-sales index » ou indice de ventes répétées

2.1. L'indice médian

L'idée la plus basique pour définir un indice de prix consiste, à partir des transactions réalisées à une date donnée, à calculer le prix au m² pour chacune d'elle et à prendre la moyenne de ces valeurs; on parle dans ce cas d'indice médian. Pour appliquer cette méthode il faut pouvoir disposer à chaque date d'un nombre suffisant de données afin d'assurer une certaine stabilité de la moyenne. Or, pour certains types de biens comme l'immobilier commercial ou de bureaux, les échantillons sont parfois de petite taille. De plus, effectuer une moyenne brute sur des biens hétérogènes, c'est courir le risque de capturer les variations dans la qualité des biens (rénovation, agrandissement…), ou de capturer les variations des prix implicites des caractéristiques de ces biens (engouement passager pour les appartements avec terrasse…), plutôt que d'essayer de retranscrire les fluctuations de la valeur foncière intrinsèque. Le problème de la rareté des données, dû à la faible liquidité du marché immobilier, se traite en général en ayant recours à des expertises. Les difficultés induites par l'hétérogénéité sont à l'origine des méthodes hédoniques.

2.2. L'indice sur valeurs d'expertise

Les indices immobiliers sont en général publiés sur une base semestrielle ou trimestrielle. Or, dans une perspective d'investissement, ces données basse fréquence sont parfois insuffisantes. Le passage à un rythme mensuel est cependant difficile car se pose alors la question du seuil critique de transaction à collecter chaque mois pour atteindre un niveau satisfaisant de fiabilité de l'indice. L'exemple de l'immobilier de bureaux est symptomatique de cette situation où le souhait des investisseurs de disposer de benchmarks réactifs bute sur la rareté des données. La solution consiste alors à construire des indices en se servant non pas des données de transactions mais de valeurs d'expertise (ou éventuellement des deux sources comme dans certaines méthodes hybrides). On considère en général qu'une expertise est convenable si

l'erreur est inférieure à 10% du prix réel. Si l'on suppose que cet exercice n'est pas affecté de biais, en agrégeant les données on obtient alors, par compensation des erreurs entre elles[6], une valeur moyenne assez fiable. En Europe ce genre d'indice est par exemple développé par IPD, Investment Property Database (indices consultables en ligne à www.ipdindex.co.uk). En passant des prix bruts des transactions aux valeurs issues des processus d'évaluation réalisés par les experts, on passe d'une situation de données objectives à une situation de données subjectives, et ce changement de paradigme n'est pas sans conséquences. La réflexion scientifique sur ce thème s'est développée autour de deux axes : le lissage et l'autocorrélation (au niveau des biens et au niveau de l'indice).

La première direction de recherche étudie le fait que l'expert ne donne qu'une valeur moyenne. Se fonder sur des évaluations revient alors à sous-estimer la dispersion des prix. Si cette caractéristique se transfère au niveau de l'indice, la volatilité calculée sera plus faible que la volatilité réelle et l'indice ne rendra donc pas compte correctement du risque de l'investissement immobilier. Pour remédier à ce problème, on essaie alors de mettre en place des procédures de délissage de l'indice pour lui redonner son vrai niveau de risque, cf. Geltner (1991). Le débat n'est cependant pas tranché sur ce thème, Lai et Wang (1998) affirmant par exemple que le lissage ne passe pas des biens à l'indice et que la volatilité de ce dernier est correcte.

L'autre thème de recherche analyse les effets de l'autocorrélation induite par l'expertise. Si un évaluateur observe dans le marché une transaction réelle accidentellement élevée, il sera amené à réviser ses croyances à la hausse et ses prochaines expertises seront alors trop élevées. Le bruit créé par cette transaction non représentative se propagera aux dates postérieures avant d'être neutralisé par les transactions ultérieures (force de rappel du marché). Au niveau de l'indice, ce phénomène peut engendrer des séries indicielles auto-corréelées, cf. Brown, Matysak (2000).

[6] Loi des grands nombres

2.3. L'indice hédonique

L'approche hédonique consiste à retrancher au prix d'un bien les valeurs implicites de ses caractéristiques (étage, nombre de salles de bains, qualité de l'environnement urbain…) pour obtenir une valeur foncière intrinsèque ; on élimine ainsi le problème de l'hétérogénéité. La décomposition hédonique du prix s'écrit :

$$P = \delta + \sum_{k=1,\dots K} \beta_k X_k + \varepsilon$$

Où : - P est le prix du bien
- X_k la valeur de la k^{eme} caractéristique
- ε l'erreur du modèle : espérance nulle, variance constante

Cette régression est effectuée à chaque instant t en se fondant sur les transactions observées à cette même date ; on détermine ainsi les valeurs des K + 1 estimateurs ($\hat{\delta}_t$, $\hat{\beta}_{1t}$, … , $\hat{\beta}_{Kt}$). L'indice est ensuite défini comme l'évolution des prix d'un bien standard (par exemple le pavillon de banlieue à trois chambres et deux salles de bains pourvu d'un terrain de 300m²) dont les caractéristiques (X_1^*, … , X_1^*) sont supposées constantes dans le temps. En d'autres termes il se calcule par :

$$I_t = \hat{\delta}_t + \sum_{k=1,\dots K} \hat{\beta}_{kt} X_k^*$$

Cette technique permet d'obtenir des indices de prix mais également de construire des systèmes d'expertise automatisés fournissant un prix, ou plus exactement une fourchette de prix, à partir d'une série de caractéristiques.

Comme indiqué dans Hoesli, Giaccotto, Favarger (1997), dont les notations introduites ci-dessus sont issues, on peut aussi construire un indice hédonique en

effectuant une unique régression lorsque les échantillons d'estimation sont de taille réduite :

$$P = \sum_{t=1,\ldots,T} \delta_t D_t + \sum_{k=1,\ldots K} \beta_k X_k + \varepsilon$$

Les D_t sont des variables muettes (« dummies ») indiquant la date d'achat du bien ; les prix implicites des caractéristiques[7] sont supposés constants dans le temps. L'indice immobilier est alors défini par :

$$I_t = \hat{\delta}_t + \sum_{k=1,\ldots K} \hat{\beta}_k X_k^* \qquad \text{ou plus simplement par}^{8} : \quad I_t = \hat{\delta}_t$$

La modélisation hédonique amène deux remarques. En premier lieu, elle suppose implicitement que la forme de la relation théorique entre le prix et les caractéristiques est connue. Les formules présentées ci-dessus sont en effet linéaires, or on pourrait tout à fait imaginer un modèle logarithmique ou d'un tout autre type. Bao et Wan (2004) utilisent par exemple une modélisation semi-paramétrique dans laquelle les impacts de certaines caractéristiques sont modélisés grâce à des fonctions splines[9].

L'autre point important pour l'emploi de ces méthodes est la définition des caractéristiques. Ce choix est de la responsabilité de l'économètre et il n'existe pas de sélection universellement reconnue. On utilise, classiquement, une dizaine de variables en veillant à éviter les problèmes de multicolinéarité. Hormis les regresseurs classiques décrivant physiquement le bien (étage, surface, nombre de salles de bains, âge de l'immeuble[10]…) il est possible d'introduire beaucoup d'autres éléments, en fonction de l'objectif de l'étude que l'on souhaite réaliser. Tong et

[7] les β_k

[8] La somme $\sum_{k=1,\ldots K} \hat{\beta}_k X_k^*$ étant constante dans le temps elle n'apporte pas d'information sur l'évolution des prix, on peut donc la retrancher

[9] Les splines sont des fonctions polynomiales par morceaux ayant de bonnes propriétés de recollement. Le but de cette approche semi-paramétrique est d'autoriser une grande flexibilité pour la relation théorique en ne la contraignant pas à suivre une structure prédéterminée trop rigide.

[10] Le regresseur « âge de l'immeuble » occupe en général une place particulière, cf. Clapp, Giaccotto (1998)

Glascock (2000) analysent ainsi l'impact de la structure d'habitation (copropriétés, maisons mitoyennes, maisons isolées). Wolverton, Senteza (2000) pointent l'importance d'une analyse régionale des indices. Francke et Vos (2004) étudient les effets croisés de la région et du type de biens. Pour mener une étude économique Engberg, Greenbaum (1999) introduisent une variable dummy destinée à capturer l'impact des zones franches sur les prix immobiliers des biens alentours. Le champ de l'économétrie spatiale, actuellement en plein développement, peut également produire des regresseurs hédoniques pertinents : Tu, Yu, Sun (2004), Berg (2005) ou Clapp (2004). Cette liste succincte n'est bien sûr pas exhaustive. Pour un exemple d'application sur données françaises on pourra consulter Beauvois, David et al. (2005) ou Laferrère (2004).

Il faut noter qu'il est également possible d'appliquer la méthode hédonique aux loyers, cf. Hoesli, Thion, Watkins (1997). Enfin, si cette technique est plutôt employée pour l'immobilier résidentiel, elle est également utilisable pour l'immobilier commercial, cf. Munneke, Slade (2001).

2.4. L'indice de ventes répétées

Les trois méthodologies présentées ci-dessus mettent la notion de prix au centre de l'estimation. Intuitivement, la valeur de l'indice à la date t est une moyenne sophistiquée des transactions (ou des valeurs d'expertises) réalisées à l'instant t. Pour l'approche hédonique, les hypothèses économiques sous-tendant le modèle sont loin d'être anodines. On affirme qu'il existe des marchés implicites pour les caractéristiques permettant de leur attribuer des prix. On fait d'autre part une hypothèse de rationalité très forte sur les agents économiques en supposant qu'ils évaluent leur bien en appliquant un processus de décomposition du tout en ses parties. De plus, en passant d'une donnée brute (le prix de la transaction) à une valeur abstraite (le δ des régressions hédoniques) on effectue un saut épistémologique

puisqu'on ne travaille plus alors sur la manifestation phénoménale elle-même mais sur une retranscription de celle-ci. Sur un autre plan, la définition des caractéristiques peut se révéler dangereuse, car cet exercice crée un risque de voir les efforts de recherche dériver vers une zoologie énumérative des éléments constitutifs des biens, plutôt que d'essayer d'introduire des concepts synthétiques en lien avec une démarche fondamentale.

La méthode des ventes répétées est à la fois moins ambitieuse scientifiquement que la technique hédonique, car on ne cherche pas à expliquer pourquoi le prix est le prix, mais également plus orthodoxe car la distance du modèle aux faits est plus faible (le lien se fait sans l'intermédiation d'une décomposition présupposée de la valeur). La notion au centre de cette théorie n'est plus le prix mais le rendement ; on se place donc dans une optique très financière. La relation fondamentale servant à construire l'indice s'écrit :

$$Ln(p_{k,j} / p_{k,i}) = ln(Indice_j / Indice_i) + \varepsilon$$

Un bien k a été acheté à la date i au prix $p_{k,i}$ et revendu à la date j au prix $p_{k,j}$, générant un rendement (continu) égal à $Ln(p_{k,j} / p_{k,i})$. L'estimation de l'indice se fait en affirmant que le rendement réalisé sur ce bien est égal au rendement réalisé sur l'indice entre les mêmes dates plus un terme d'erreur. A la différence des indices de prix (médian, hédonique…) l'échantillon d'estimation[11] d'un indice de ventes répétées est constitué de biens pour lesquels on connaît deux valeurs consécutives de transaction. On n'entrera pas ici plus en détail dans cette modélisation car elle sera largement développée tout au long de cette thèse.

[11] Ce type d'échantillon est en général plus petit que ceux utilisés pour les indices hédoniques, car observer une vente répétée est moins fréquent qu'observer une vente simple. Cette méthode est également plus légère car elle ne requiert pas l'information très précise et très volumineuse nécessaire au fonctionnement de la technique hédonique : la capture des variations de la qualité suppose en effet une description détaillée de chaque bien.

Les variations de qualité des biens ne sont pas intégrées dans cette approche. Pour bien comprendre ce que capture cet indice prenons l'exemple d'un propriétaire qui décide d'aménager des combles en investissant dans sa maison une certaine somme pour réaliser les travaux, augmentant ainsi la qualité de son logement et donc sa valeur. A la date de revente, si on rapporte le prix d'achat au prix de vente, on calcule la rentabilité brute de l'actif immobilier. Pour obtenir une rentabilité nette il faudrait retrancher au prix de revente la valeur capitalisée des investissements consentis (les travaux). L'indice de ventes répétées est donc un indice financier estimant le rendement brut de l'immobilier. En d'autres termes et au risque d'être caricatural : "Le prix c'est le prix, le rendement c'est le rendement". Peu importe pourquoi le prix est à ce niveau, la seule chose qui compte comme objet d'étude c'est la valeur de transaction. Si un propriétaire a réalisé des travaux il lui appartiendra d'intégrer les coûts afférents pour calculer la rentabilité nette et de mener plus en détail l'analyse du bilan de son opération. Intégrer cette notion complémentaire à un indice ne pourrait d'ailleurs que donner une valeur moyenne bien moins pertinente que celle issue d'une analyse au cas par cas.

Les deux articles ayant popularisé la technique des ventes répétées sont Bailey, Muth, Nourse (1963) et Case, Shiller (1987). Il faut également mentionner les travaux réalisés antérieurement par G.Duon sur données françaises Duon (1943a-1943b-1946) et prolongés par J. Friggit (2001) retraçant l'évolution des prix de l'immobilier résidentiel depuis 1840.

2.5. Les indices hybrides

L'existence d'approches indicielles concurrentes pose le problème du choix de LA bonne méthode. Or, bien que ce thème ait été débattu vigoureusement depuis 20 ans, aucune technique n'a encore vraiment supplanté les autres. On ne rentrera pas ici dans une étude détaillée des avantages et des inconvénients respectifs de ces

méthodes ; on pourra se reporter sur ce point à Wang, Zorn (1999) pour avoir un éclairage intéressant sur ce qui n'est, peut-être, qu'un faux débat.

Si la comparaison des méthodes est en général productive dans le sens où elle permet de mieux saisir la pertinence de chacune d'entre elles, on peut également envisager de les combiner pour obtenir ce que l'on appelle des indices hybrides. Cette idée est apparue très tôt dans la littérature. Ainsi un des premiers articles sur ce sujet, Case, Quigley (1991) a été publié seulement quatre ans après l'article fondateur de Case, Shiller (1987). Le principe de cette méthode consiste à effectuer une estimation jointe d'un modèle hédonique, sur l'ensemble des biens pour lesquels on ne dispose que d'un prix, et d'un modèle de ventes répétées pour ceux dont on connaît deux prix de transactions[12]. L'article de Case, Pollakowski, Watcher (1991) discute et évalue l'apport de cette approche par rapport aux techniques de base ; Quigley (1995) et Englund, Quigley, Redfearn (1998) sont des approfondissements de cette réflexion.
L'hypothèse implicite aux indices de ventes répétées est la constance de la qualité du bien entre l'achat et la revente. Or, si pour la majorité des caractéristiques cette hypothèse peut sembler raisonnable en l'absence de travaux de rénovation, il existe cependant des variables qui, par nature, ne peuvent pas être considérées constantes ; l'age du bien en est l'exemple typique. Il n'existe donc pas de données de ventes répétées à qualité parfaitement inchangée et, pour éviter les biais, il peut être utile d'incorporer une certaine dose d'approche hédonique à la spécification traditionnelle de l'indice Case, Shiller. On obtient alors une autre méthode hybride développée par Cannaday, Munneke, Yang (2005) et Chau, Wong, Yiu (2005).
Enfin on peut mentionner l'article de Clapp, Giaccotto (1992), où les trois techniques basiques (expertises, ventes répétées, décompositions hédoniques) sont utilisées conjointement, et l'article de Geltner, Goetzmann (2000)[13] qui croise l'approche des ventes répétées avec des valeurs d'expertise. Il existe donc, comme on le constate, une grande variété de techniques hybrides.

[12] On distinguera d'ailleurs les ventes répétées à qualité inchangée des ventes répétées à qualité modifiée
[13] repris par Hordijk, De Kroon, Theebe (2004).

3. Indices immobiliers, gestion de portefeuille et produits dérivés sur indices immobiliers

Au cours des dernières années l'immobilier a émergé comme classe d'actifs à part entière. De nombreuses recherches ont été, et seront, menées pour déterminer sa juste place dans les portefeuilles des investisseurs, comme en témoignent les états de l'art programmatiques réalisés par Dombrow, Turnbull (2004) et Newell et al. (2004). Ce support d'investissement a l'avantage de combiner des rendements corrects et relativement stables avec un grand potentiel de diversification, sa corrélation avec les actions ou les obligations étant assez faible. Selon le profil de l'investisseur, intégrer de l'immobilier dans un portefeuille mixte permettra donc de réduire le risque en conservant la même rentabilité, ou d'augmenter la rentabilité pour un même niveau de volatilité.

L'investissement peut se faire en utilisant des actions de foncières cotées[14], ou grâce à l'achat de parts de SCPI (Société Civile de Placement Immobilier) dont le statut vient d'être récemment réformé pour en faciliter la gestion (on parlera dorénavant d'OCPI : Organisme collectif de placement immobilier). On pourra se reporter à Lee, Stevenson (2005) ou Schoeffler (2005) pour mesurer l'apport de ces titres aux portefeuilles traditionnels[15]. Pour ce genre de placement on parle « d'immobilier indirect » car l'investisseur n'achète pas directement les biens, mais il prend simplement des participations dans une structure chargée de le faire. Dans une optique de recherche, ces titres présentent le grand avantage d'être cotés ou expertisés, plus ou moins régulièrement. Il est donc possible de mettre en place des études de type Markowitz.

[14] En France, le terme SIIC (Société d'Investissement Immobilier Cotée) est aussi employé pour parler des foncières. Ce sigle désigne plus particulièrement celles qui choisissent d'adopter un régime fiscal spécifique. Plus généralement au niveau international on parlera de REIT (Real Estate Investment Trust).

[15] Le premier article étudie l'intérêt des REIT en fonction de l'horizon d'investissement, le deuxième est une étude du cas français.

Si l'on se tourne vers la détention « d'immobilier direct » (sans intermédiation d'une structure financière), la problématique de gestion de portefeuille devient plus délicate car cet actif n'est pas coté. Les cours des actions de foncières peuvent éventuellement servir de proxies mais il faut alors les manier prudemment car, jusqu'à une date récente, ils présentaient une corrélation non négligeable avec les indices boursiers classiques. Les utiliser comme substituts fait donc courir le risque de prendre des décisions sous-optimales, car perturbées pas un risque de marché qui n'a pas sa place. Il existe des tentatives pour essayer de neutraliser la composante "portefeuille de marché" des SIIC[16], cf. Stevenson (2000), mais l'idéal reste cependant de disposer de bases de données conséquentes et de bonne qualité sur l'immobilier direct. A partir de ces bases, on peut construire des indices par secteur, par région, par type de biens et les intégrer ensuite aux études de choix de portefeuille. Hoesli, Lekander, Witkiewicz (2005) et Lee (2005) étudient par exemple, grâce aux indices, les gains générés par la diversification immobilière. Sur ce thème Kaiser (2004) va même jusqu'à considérer que les obligations sont supplantées par l'investissement immobilier.

Cette approche indicielle peut être approfondie en utilisant différents concepts en provenance de la finance des actions. Marcato (2004) propose par exemple de créer des indices "growth" et "value" permettant de mesurer les performances des propriétés en mettant l'accent soit sur les loyers perçus (value), soit sur les plus-values à la revente (growth). Dans le même ordre d'idée, Young (2005) propose de modifier légèrement la méthodologie traditionnelle de calcul de l'indice américain NCREIF pour le rendre plus lisible par les investisseurs, et faciliter ainsi les comparaisons avec les autres classes d'actifs. Ces deux derniers articles témoignent particulièrement bien de la financiarisation de l'immobilier actuellement en cours et de l'intérêt de développer une théorie des indices immobiliers performante pour pouvoir l'appliquer aux problématiques de gestion de portefeuilles.

[16] SIIC : Société d'Investissement Immobilier Cotée (cf. note 14).

Le raisonnement et la pratique financière peuvent même être poussés plus avant en donnant à l'indice le statut d'actif négociable. On ne le considère plus alors comme un simple benchmark permettant de mesurer les performances d'un placement. Des produits dérivés sur indice immobilier ont ainsi été récemment introduits en Angleterre et aux Etats-Unis pour faciliter la gestion du risque immobilier : il est désormais possible d'acheter ou de vendre l'indice. Ce genre de produits, essentiellement des swaps et des futures, semble rencontrer un succès significatif auprès des investisseurs (1 milliard d'euros d'encours pour l'année 2005 à Londres).

Des particuliers pourraient même y avoir recours, directement ou indirectement, pour gérer leur patrimoine. L'achat d'un bien immobilier constitue en effet la plus grande part du portefeuille des ménages, et il n'est pas rare que cet actif représente 300% de la richesse nette en raison de sa non-divisibilité. L'aspect "tout ou rien" de ce support surexpose donc inévitablement les ménages au risque immobilier, à des niveaux qu'ils ne souhaitent probablement pas[17]. L'introduction de produits dérivés sur indices permettrait d'une part de répondre au problème d'indivisibilité et, dans le même temps, de rendre possible les positions courtes. Un ménage pourrait alors décider de se couvrir, totalement ou partiellement de ce risque, tout en restant propriétaire. Ces problématiques ont été étudiées dans un premier temps par Flavin, Yamashita (2002) pour les Etats-Unis puis par Englund, Hwang, Quigley (2002) pour la Suède, Iacoviello, Ortalo-Magné (2003) pour l'Angleterre et Le Blanc, Lagarenne. (2004) pour la France.

Pour être performant, de tels produits dérivés doivent être construits sur un indice de bonne qualité, suffisamment représentatif du marché concerné. De plus, l'évaluation de ces titres ne pourra être mené efficacement que si les caractéristiques de l'actif sous-jacent sont connues et bien comprises. Les indices se voient donc légitimés comme sujet de recherche à part entière. On pourra consulter Buttimer, Kau, Slawson

[17] Cette situation a d'ailleurs d'importantes implications pour les études cherchant à expliquer la détention de titres financiers classiques (actions et obligations) par les ménages. Ne pas prendre en compte l'immobilier c'est courir le risque d'arriver à des résultats erronés, car le niveau de risque auquel doit faire face le particulier est alors appréhendé improprement.

(1997) ou Björk, Clapham (2002) pour avoir des éléments sur l'évaluation de ces dérivés, mais il faudra bien garder à l'esprit que la théorie de l'arbitrage devra être maniée avec précaution dans le domaine de la finance immobilière, comme on le verra dans le chapitre 4, paragraphe 4.

4. Indices et prêts immobiliers

Lorsqu'une banque accorde un prêt immobilier à un client elle espère avoir récupéré, au terme du contrat et selon l'échéancier prévu, l'intégralité du capital prêté plus une rémunération pour la location d'argent. Comme dans tout contrat de prêt, il existe un risque que l'emprunteur ne puisse pas, ou ne veuille pas, honorer ses engagements et qu'en conséquence le prêteur subisse une perte sur le capital ou les intérêts. La modélisation du risque de défaut fait actuellement l'objet de nombreuses recherches, souvent très pointues. Sans entrer dans des considérations techniques sophistiquées, on décompose habituellement l'espérance mathématique de la perte L en trois composantes (cf. Bluhm, Overbeck, Wagner 2003, page 17 pour plus de détails) :

$$E[L] = EAD \times LGD \times P(D)$$

Où :

- EAD mesure l'exposition au risque, en euros ("Exposure At Défault")
- LGD mesure la perte en cas de défaut, en pourcentage ("Loss Given Default")
- P(D) correspond à la probabilité du défaut

Dans les modèles de risque de crédit, la probabilité de défaut est en général difficile à estimer et cela est particulièrement vrai pour les prêts immobiliers. La détermination de cette grandeur doit en effet faire intervenir des considérations macroéconomiques (niveau des salaires, chômage, retraite…), des considérations

individuelles (divorce, accidents de la vie…), voire des considérations stratégiques[18]. A l'opposé, l'exposition au défaut (EAD) est très simple à calculer puisqu'il s'agit simplement du capital restant dû à une date donnée. Si ces deux grandeurs n'ont que peu de rapport avec les prix de l'immobilier et donc avec les indices, il en va tout autrement pour la « LGD » comme nous allons le voir.

Un prêt immobilier est toujours couplé à une garantie. Celle-ci peut prendre la forme d'une hypothèque ou d'une caution[19]. Dans le cas d'une hypothèque, lorsque l'emprunteur fait défaut, le bien est vendu (ou saisi) et deux cas sont alors possibles. Soit le produit de la vente permet à la banque de recouvrer l'intégralité de la dette, elle ne supporte alors pas de pertes (hormis les frais juridiques). Soit la somme récupérée est insuffisante et le prêteur subit une perte égale à la différence entre le montant de la dette et le prix de revente du bien. Ce cas se produit en général dans un contexte de chute des prix immobiliers et concerne surtout les prêts pour lesquels la part de l'apport personnel est faible.[20] Lorsqu'un emprunteur se retrouve à une date donnée dans une configuration où la valeur de marché de son bien est inférieure à sa dette, on parle de situation de « negative equity », ou de « fonds propres négatifs » (cf. ANIL 1999). Pour une banque ces clients sont particulièrement à risque car, en cas de défaut, leur ratio LGD sera supérieur à 0 et il y aura une perte nette[21]. Cette troisième composante de l'espérance E[L] dépend donc directement des prix immobiliers.

[18] Aux Etats-Unis faire défaut dans un mortgage (prêt immobilier hypothécaire) est parfois une option stratégique intéressante financièrement, cf. Lacour-Little, Malpezzi (2003).

[19] Dans la plupart des pays le système de l'hypothèque est prépondérant. Le cas de la France, où des sociétés comme Crédit Logement cautionnent une part non négligeable des prêts immobiliers, est une exception (cf. Baude, Bosvieux 2002)

[20] Ou de manière équivalente avec un niveau élevé de LTV. Le "loan to value" est égal au rapport, à une date donnée, entre le capital restant du et le prix de marché du bien immobilier.

[21] Il faut remarquer ici que ce concept de « negative equity » ne concerne pas seulement les particuliers qui achètent pour se loger. Il est aussi pertinent pour les propriétaires bailleurs qui utilisent des montages avec peu de fonds propres pour profiter d'un effet de levier important (Batsch 2005). Le mécanisme est identique : si la valeur du bien s'effondre peu de temps après l'origination du prêt, le risque apparaît pour l'établissement de crédit.

Si le prêt est assorti d'une caution, la banque est couverte intégralement contre le risque de défaut car l'organisme de caution prend alors le relais de l'emprunteur défaillant. Il s'agit en fait d'un mécanisme d'assurance : les primes versées en début de prêt par l'ensemble des emprunteurs servent à compenser les pertes éventuelles de certains d'entre eux.

Il faut noter, de plus, que dans le cas des prêts hypothécaires la banque peut choisir de conserver ce risque à son bilan et le gérer directement (Demey, Frachot, Riboulet 2003 ou Simon 2005), ou bien l'externaliser en utilisant différents véhicules de refinancement. Pour ce faire elle peut créer un fonds commun de créances[22], une société de crédit foncier[23], ou encore utiliser des véhicules assimilables aux sociétés de crédit foncier sans en avoir pour autant le titre[24]. Il existe bien sur des différences entre ces techniques, mais l'idée centrale et commune repose sur la titrisation. Le risque associé au défaut d'un prêt immobilier n'est plus supporté par la banque ni par un organisme d'assurance mais par les investisseurs qui achètent les titres obligataires émis par ces véhicules, en représentation des prêts. La tension qui apparaît sur les cash-flows en cas de "negative equity" ou de défaut effectif est alors gérée grâce aux procédures mises en place dans la structure (overcollatéralisation, subordination, réserves de liquidité …) et, en dernier recours, les pertes seront supportées par les investisseurs[25].

En résumé, un prêt comporte donc toujours une exposition indirecte au risque immobilier, via le ratio LGD. Ce risque est irréductible mais il peut être transféré des banques aux sociétés de caution ou aux investisseurs. Quelque soit le maillon de la chaîne financière qui le supporte, la mesure correcte des prix immobiliers, grâce aux techniques indicielles, est incontournable pour pouvoir réaliser une gestion correcte.

[22] Titrilog 1998 pour Calyon par exemple (cf. bibliographie)

[23] CIF Euromortgage pour le Crédit Immobilier de France (cf. Caisse centrale du crédit immobilier de France, CIF Euromortgage ou CIFD dans la bibliographie) ou la Compagnie de Financement Foncier pour le Crédit Foncier de France (cf. bibliographie)

[24] Vauban Mobilisations Garanties (cf. bibliographie) et la Caisse de Refinancement de l'Habitat (cf. CRH dans la bibliographie)

[25] Les investisseurs sont en général également confrontés au risque de remboursement anticipé ; on pourra consulter sur ce thème Frachot, Gourieroux (1995)

Enfin, il faut noter que vis-à-vis des produits dérivés, ces différents acteurs financiers constituent des vendeurs naturels de risque immobilier.

5. Marché immobilier et enjeux sociaux

Si on peut étudier le marché de l'immobilier résidentiel sous un angle strictement financier, en posant par exemple la question de son efficience comme dans Case, Shiller (1989), d'autres problématiques plus économiques sont également légitimes.

Ainsi, dans la période actuelle qui se caractérise par des prix immobiliers élevés dus à une très forte croissance sur les dix dernières années, il peut être relativement difficile d'acheter pour des primo-accédants ou pour des familles modestes. Même si les organismes de financement se sont adaptés en augmentant sensiblement la durée des crédits immobiliers, cf. Bosvieux, Vorms (2003), la situation actuelle est assez inédite. Afin d'essayer de comprendre ce phénomène, que certains qualifient de bulle, on peut essayer de le replacer dans une perspective historique en étudiant par exemple les évolutions du ratio (prix des logements / revenu disponibles). On pourra se reporter sur ce thème à Friggit (2002) ou consulter le site de l'adef (http://www.foncier.org/statistiques/accueil_statistiques.htm).
On peut d'autre part mentionner les problématiques liées à l'aménagement du territoire. Si les prix immobiliers sont trop élevés, ils peuvent être à l'origine d'une segmentation urbaine et sociale où les familles les moins favorisées se retrouveront dans les mêmes zones (ghettoïsation). Dans le même ordre d'idée, l'envolée des prix constatée dans certaines zones touristiques peut conduire à une certaine éviction des habitants du crû. On pourra par exemple penser à la situation surprenante du paysan de l'île de Ré, dont les revenus sont proches du SMIC, mais pourtant soumis à l'ISF en raison de la très forte augmentation de la valeur foncière de ses terrains sous la pression touristique.

Accession à la propriété, surendettement, bulle immobilière, politique de la ville, aménagement du territoire… Pour pouvoir traiter tous ces thèmes correctement, il est nécessaire de savoir construire des indicateurs fiables et dont les rouages théoriques sont bien compris. La notion d'indice immobilier est donc, ici aussi, centrale. Mais plus largement il en est de même de toutes les techniques d'analyse de données immobilières issues des constructions indicielles, car plus le diagnostic sera précis, plus il sera possible de cibler les points critiques sur lesquels devront se porter les efforts de politique urbaine. On mettra par exemple en œuvre dans le chapitre 2, paragraphe 3, une méthodologie qui permettra d'avoir une vision assez fine des comportements des propriétaires (incitations à l'achat, incitations à la revente, durée de détention des biens…).

6. Plan général et articulation des chapitres

De manière synthétique, cette thèse propose une reformulation et un approfondissement théoriques de l'indice de ventes répétées. Les conséquences et les applications pratiques découlant de ce nouveau formalisme viendront légitimer la démarche initiale qui pourra, dans un premier temps, paraître un peu abstraite. Les différents chapitres s'organisent et s'articulent comme suit.

Ce chapitre introductif a décrit le contexte général dans lequel s'inscrit cette recherche et a évoqué quelques problématiques classiques de la finance immobilière. Les difficultés théoriques du concept de prix de marché de l'immobilier ont été présentées ainsi que les différentes méthodologies indicielles (la variété de ces dernières n'étant en fait que l'illustration empirique de la complexité de la notion de mesure dans les marchés hétérogènes et illiquides). La gestion de portefeuille, les produits dérivés, les prêts immobiliers, les enjeux politiques et sociaux ont également été évoqués. L'objectif de cette partie consistait à mettre en évidence les enjeux et à

tenter de convaincre le lecteur de l'importance de développer une réflexion poussée, en Sciences de Gestion, dans ce champ d'application particulier qu'est l'immobilier. Jusqu'à une période récente ce secteur n'était en effet pas perçu comme relevant de la finance et comme étant susceptible d'être le support de procédures d'optimisation économique. Cette période semblant maintenant révolue, la recherche financière en France devrait probablement trouver un intérêt grandissant à explorer cette Terra Incognita, dans laquelle beaucoup de choses restent à inventer, tout en s'appuyant sur des méthodologies et des réflexions initiées dans d'autres pays ayant une tradition plus ancienne dans ce domaine.

Le premier chapitre débutera avec une question un peu formelle : Peut-on établir un lien fonctionnel entre un indice de prix et un indice de rendement comme le RSI ? Pour répondre à cette interrogation, nous serons amenés à réexaminer en détail la structure de l'indice de ventes répétées et à le réécrire en introduisant différents concepts intermédiaires. Ces notions présenteront le grand avantage d'être interprétables financièrement et faciles à manier. En s'appuyant sur la décomposition qui aura émergé à cette occasion, nous construirons conjointement un échantillon de référence pour l'étude des ventes répétées, basé sur une modélisation exponentielle de la décision de revente.
Le premier temps de ce chapitre sera d'ordre inductif : nous chercherons à répondre à la question initiale, dans un cadre simplifié. Par la suite, le nouveau formalisme pour le RSI sera présenté en toute généralité, dans une démarche plus déductive. La réécriture théorique de l'indice de ventes répétées à laquelle nous aboutirons constituera le cœur de cette thèse.

Le deuxième chapitre explorera les premières conséquences empiriques de la reformulation théorique. Nous y présenterons une méthodologie d'analyse des données immobilières de ventes répétées qui, grâce à divers indicateurs, assurera une exploitation de l'information enchâssée dans les échantillons bien supérieure à celle que l'on aurait obtenue en se contentant simplement de calculer l'indice. Nous

réaliserons également une étude de la sensibilité du RSI aux paramètres du modèle et aux caractéristiques de la distribution de l'échantillon.

Le chapitre trois prolongera l'étude des conséquences du chapitre 1 en examinant, sous ce nouvel éclairage, plusieurs problèmes classiques de la littérature. Nous étudierons ainsi la volatilité de l'indice et le problème de la réversibilité. Les formules auxquelles nous aboutirons seront à la fois simples et intuitives, elles permettront d'appréhender facilement certaines caractéristiques un peu complexes du RSI. Le problème des deux populations sera également étudié dans cette section.

Enfin, le quatrième et dernier chapitre ouvrira des pistes d'études pour de futures recherches. Nous y développerons les premiers éléments d'une théorie des indices informationnels dans laquelle la quantification de l'information ne sera plus implicite, comme dans le modèle traditionnel de Case-Shiller, mais explicite et donc adaptable. Nous examinerons également la structure de la matrice d'information à l'aide d'outils géométriques. Finalement, la possibilité d'appliquer la théorie de l'arbitrage à l'évaluation des produits dérivés sur indices immobiliers sera discutée, tant du point de vue théorique qu'empirique.

Chapitre 1

Un nouveau formalisme pour la méthode des ventes répétées

1. Introduction

Le point de départ de ce chapitre porte sur l'existence d'une relation théorique entre les indices de prix (médian, hédonique …) et de l'indice de ventes répétées. Pour répondre à cette question le paragraphe 2 présente au préalable le détail la méthodologie du RSI. On y introduira la notion qui se révélera être centrale dans cette thèse, à savoir la quantité d'information associée à une transaction donnée, et les intuitions amenant à soupçonner l'existence d'un lien formel entre les deux indices seront discutées. Le paragraphe 3 résoudra le problème dans un environnement volontairement simplifié, afin d'éviter une trop grande pesanteur lors de l'exposition, et surtout pour familiariser le lecteur avec les différents concepts qui apparaîtront à ce propos. Car, en effet, au cours de la résolution de ce problème initial un formalisme très intéressant se dégagera, laissant entrevoir de multiples applications. Il sera donc important de bien comprendre comment les concepts introduits dans la démonstration simplifiée s'articuleront entre eux. Dans les paragraphes 4 et 5 on présentera, toujours en profitant de la flexibilité du cas élémentaire, la construction d'un échantillon benchmark de ventes répétées. On étudiera d'abord ses propriétés théoriques avant de donner un aperçu de la manière dont il pourra être employé dans une analyse de données.

Le dernier paragraphe renversera la posture de recherche en passant de la démarche exploratoire et inductive des paragraphes 2 et 3 à une reformulation déductive, et en toute généralité, de la structure fine du RSI, telle qu'elle est apparue au cours de la résolution de la question introductive. On présentera la décomposition algorithmique de l'indice en briques élémentaires interprétables, on récapitulera les notations et les résultats théoriques généraux seront établis rigoureusement. Etant donné que la structure fondamentale de ce nouveau formalisme ne sera pas modifiée par l'exercice de généralisation, les démonstrations détaillées des résultats de ce paragraphe seront présentées en annexe et pourront être ignorées en première lecture. La question de la persistance du lien fonctionnel entre les deux indices sera analysée, mais cette fois en

tant que corollaire d'une réécriture fondamentale du RSI et non plus en tant que problématique principale.

2. La problématique initiale

Comme on a déjà pu s'en apercevoir les méthodologies indicielles sont nombreuses et variées. Si l'on calcule plusieurs types d'indices à partir d'un même échantillon les résultats peuvent, dans une certaine mesure, différer[1]. On est alors amené à s'interroger rapidement sur le signal à privilégier pour une étude économique ou pour un choix d'investissement.

La littérature regorge d'articles empiriques comparant les performances des différentes méthodes. Toutefois, comme ces analyses sont la plupart du temps réalisées en s'appuyant sur des échantillons particuliers, leurs résultats doivent être maniés avec précaution et en gardant à l'esprit qu'ils sont probablement soumis à une certain degré de contingence. Au niveau théorique, la comparaison des méthodologies dans la littérature indicielle reste très largement d'ordre qualitatif ; à l'exception notable de l'article de Wang et Zorn (1997) où une démarche plus formelle et quantitative est adoptée. Le point de départ de ce chapitre s'inscrit dans cette problématique. Nous étudierons ici l'existence d'un lien fonctionnel, voire déterministe, entre les indices de prix (hédonique, médian …) et les indices de rendements (ventes répétées).

Ce problème n'est pas purement d'ordre théorique. En effet, s'il devient possible dans les prochaines années d'acheter des indices immobiliers, ou des produits dérivés sur ces indices, la question du lien fonctionnel entre ces différents supports amènera alors directement à la notion d'opportunité d'arbitrage. Les dérivés de deux méthodologies indicielles concurrentes, mais présentant cependant un certain degré

[1] Il arrive parfois que les retournements de tendances soient par exemple détectés plus rapidement par un indice plutôt qu'un autre.

de cohérence formelle, ne pourront pas être évalués indépendamment sans courir le risque de voir apparaître des inefficiences dans le marché.

2.1. La méthodologie des ventes répétées

L'indice de ventes répétées résulte de la décomposition suivante :

$$Ln(\, p_{k,t}\,) = ln(\, Indice_{\,t}\,) + G_{k,t} + N_{k,t} \qquad (1)$$

Avec :

- $p_{k,t}$: prix du $k^{ème}$ bien à la date t
- $Indice_t$: valeur théorique de l'indice (en niveau) à la date t
- $G_{k,t}$: marche aléatoire gaussienne[2] représentant les tendances spécifiques du $k^{ème}$ bien
- $N_{k,t}$: bruit blanc associé aux imperfections de marché pour le $k^{ème}$ bien

$Indice_t$ correspond à la valeur indicielle théorique qui est et restera inconnaissable. Les estimateurs de cette valeur seront notés par Ind_t, en d'autres termes on aura :

$$\widehat{Indice}_t = Ind_t$$

L'originalité de cette décomposition réside dans l'introduction d'une marche aléatoire dans le terme d'erreur. On peut l'interpréter en termes de représentativité d'un bien par rapport à l'ensemble du parc. Au fur et à mesure du passage du temps les caractéristiques d'un bien, en termes de normes ou de confort, peuvent s'écarter des standards du marché ; de nouvelles constructions plus au goût du jour étant introduites. Un bien ancien peut donc devenir, si des travaux ne sont pas réalisés,

[2] Pour une discussion de l'hypothèse de marche aléatoire on pourra se référer à Hill, Sirmans, Knight (1999) pour l'immobilier direct et à Kleiman, Payne, Sahu (2002) pour l'immobilier indirect.

marginal[3] par rapport aux exigences moyennes du marché et sa dynamique de prix pourra donc potentiellement diverger de la dynamique du prix de marché captée par l'indice. Dans le paragraphe suivant on verra que $G_{k,t}$ peut aussi permettre de capturer les effets des tendances locales. Historiquement, le modèle de ventes répétées de Bailey, Muth, Nourse (1963) ne possédait pas de marche aléatoire ; son introduction a été réalisée dans l'article de Case, Shiller (1987). On notera par la suite BMN le premier modèle et CS le second.

La suite de ce paragraphe présente les deux écritures matricielles possibles pour l'estimation du RSI (repeat sales index) et la procédure de régression en trois étapes associée. On réécrira également ce problème en termes d'optimisation et de minimisation des erreurs.

2.1.1. La première écriture matricielle

Pour une vente répétée avec une date d'achat i et une date de revente j on a :

$$Ln(\, p_{k,i}\,) = \ln(\, Indice_i\,) + G_{k,i} + N_{k,i} \quad \text{et} \quad Ln(p_{k,j}) = \ln(Indice_j) + G_{k,j} + N_{k,j}$$

Et en soustrayant :

$$Ln(\, p_{k,j} / p_{k,i}\,) = \ln(\, Indice_j / Indice_i\,) + \underbrace{(\, G_{k,j} - G_{k,i}\,) + (\, N_{k,j} - N_{k,i}\,)}_{\varepsilon_k} \qquad (2)$$

[3] $G_{k,t} < 0$: effet d'obsolescence $G_{k,t} > 0$: plus value due à un effet de cachet des logements anciens

$G_{k,t} \approx 0$: des travaux ont été réalisés pour suivre les standards du marché

Le taux de rendement du $k^{ème}$ couple est égal au rendement du RSI (repeat sales index) pendant la même période, plus les variations de la marche aléatoire G_k et du bruit blanc N_k. L'estimation de l'indice se fait donc sur les rendements et non pas sur les prix comme dans le cas des indices médians ou hédoniques. Comme on le verra dans la suite de cette thèse, raisonner en termes de taux est plus cohérent que raisonner en termes de niveaux absolus pour le RSI. Les différentes formules que l'on obtiendra s'écriront naturellement avec des taux et pas avec des prix.

Chacun des couples produisant une égalité de ce type, on peut les réécrire sous une forme matricielle :

$$Y = D*LIndice + \varepsilon \qquad (3)$$

Avec :

- Y : vecteur colonne des taux de rendements logarithmiques réalisés dans l'échantillon
- $LIndice = (\ln(Indice_1), \ldots , \ln(Indice_T))'$; $Indice_0$ est fixé à 1 (ou 100)
- ε : vecteur des erreurs
- D s'obtient à partir d'une matrice D' dont la première colonne a été tronqué pour éviter une matrice singulière[4] dans le processus d'estimation.
- Le nombre de lignes de D' est égal au nombre total de ventes répétées. Les (T+1) colonnes de cette matrice correspondent aux différentes dates possibles pour effectuer une transaction. Dans chaque ligne, -1 apparaît à la date d'achat, 1 à la date de revente et le reste n'est constitué que de 0.

2.1.2. La deuxième écriture matricielle

Au lieu de coder par l'intermédiaire de la matrice D la date d'achat et la date de revente, on peut transformer l'écriture matricielle pour mettre en évidence les taux qui s'appliquent à une période de détention donnée.

Par exemple si T = 3 le vecteur LIndice s'écrit ($\ln(\text{indice}_1)$, $\ln(\text{indice}_2)$, $\ln(\text{indice}_3)$) et si l'on note par r'_i les taux de croissance de l'indice théorique on a :

$$\begin{cases} \ln(\text{indice}_1) = r'_0 \\ \ln(\text{indice}_2) = r'_0 + r'_1 \\ \ln(\text{indice}_3) = r'_0 + r'_1 + r'_2 \end{cases}$$

Plus généralement il existe une matrice A telle que[5] :

$$\text{LIndice} = A\, R^* \qquad \text{avec } R^* = (r'_0, \ldots, r'_{T-1})' \qquad (4)$$

Comme l'inverse de A existe, on peut alors réécrire (3) sous la forme :

$$Y = (DA)\,(A^{-1}\,\text{LIndice}) + \varepsilon = (DA)\,R^* + \varepsilon \qquad (5)$$

Les règles basiques d'algèbre linéaire indiquent que la matrice DA a autant de lignes qu'il y a de ventes répétées dans l'échantillon et que les colonnes correspondent aux intervalles de temps élémentaires. Pour chaque ligne de DA, si l'achat survient à t_i et la revente à t_j on aura :

$$\begin{array}{ccccccccccccccc} (& 0 & 0 & \ldots & 0 & 1 & 1 & \ldots & 1 & 1 & 0 & \ldots & 0 &) \\ & 1 & 2 & & t_i\text{-}1 & t_i & t_i\text{+}1 & & t_j\text{-}2 & t_j\text{-}1 & t_j & & T & \end{array}$$

On met ainsi en avant les taux qui s'appliquent au cours des périodes de détention, car ce qu'exprime la relation (5) n'est rien d'autre que[6] :

[4] Cf. Baroni, Bathélémy, Mokrane (2004) pour les détails

[5] A est une matrice triangulaire dont les valeurs sont égales à 1 sur la diagonale principale et en dessous, 0 ailleurs.

[6] On rappelle ici que le taux r_{j-1} porte sur l'intervalle $[t_{j-1}, t_j]$

$$\text{Ln (rendement pour un couple } (t_i , t_j)) = r'_i + \ldots + r'_{j-1} + \varepsilon$$

Sur le plan des notations, on distinguera le vecteur théorique $R^* = (r'_0, \ldots, r'_{T-1})$, du vecteur estimé R. Celui-ci regroupera les estimateurs des taux continus pour les intervalles de temps élémentaires [t ; t+1] et le lien avec l'indice estimé Ind s'écrira :

$$Ind_t = \exp(r_0 + r_1 + \ldots + r_{t-1}) \qquad \text{ou} \qquad r_t = \ln(Ind_{t+1} / Ind_t)$$

2.1.3. Procédure d'estimation en trois étapes et problème de minimisation

L'estimation de la relation (3), ou de la relation (5), s'effectue en trois étapes en raison de l'hétéroscédasticité de ε. En effet la spécification du terme d'erreur dans la formule (2) produit une variance du résidu non constante :

$$Var(\varepsilon_k) = 2\,\sigma_N^2 + \sigma_G^2 (t_j(k) - t_i(k)) \qquad (6)$$

Les valeurs σ_G et σ_N sont les volatilités associées aux marches aléatoires $G_{k,t}$ et aux bruits blancs $N_{k,t}$ et l'expression $t_j(k) - t_i(k)$ correspond simplement à la longueur de la période de détention du k^{eme} bien.

Une première estimation par MCO[7] est effectuée, produisant une série de résidus. Le carré de ces résidus est ensuite régressé sur une constante et sur la longueur des périodes de détention correspondantes. On obtient ainsi un estimateur $\sum$ de la matrice de variance-covariance[8] des ε tel que :

$$\sum_{k,k} = 2\,\sigma_N^2 + \sigma_G^2 (t_j(k) - t_i(k)) \qquad \sum_{k,l} = 0 \qquad \text{si } k \neq l$$

[7] Moindres carrés ordinaires

La troisième étape consiste alors à appliquer la procédure des moindres carrés pondérés à l'équation (3) en utilisant la matrice $\sum$. Ce qui signifie, en d'autres termes, qu'il faut résoudre le problème d'optimisation suivant :

$$\text{Min}_{\text{Ind}} \left[(Y - D\ \text{Ind})' \sum^{-1} (Y - D\ \text{Ind}) \right] \qquad (7)$$

Ou encore : $$\text{Min}_{R} \left[(Y - (DA)\ R)' \sum^{-1} (Y - (DA)\ R) \right] \qquad (7')$$

Si l'on décrit les différentes possibilités pour les couples (t_i, t_j) par une première somme et que l'on répertorie par une seconde somme les transactions associées à un même couple de dates, le problème (7') se réécrit alors :

$$\text{Min}_{R} \left[\sum_{i<j} \sum_{k'} \{\ln(p_{k',j} / p_{k',i}) - (r_i + \ldots + r_{j-1})\}^2 / \{(\sigma_G^2(j - i) + 2\sigma_N^2)\} \right] \qquad (7'')$$

2.2. Mesures de bruit et d'information

Si l'on était dans un contexte de MCO, le terme multiplicatif $(2\sigma_N^2 + \sigma_G^2(j-i))^{-1}$ serait constant et il pourrait être sorti du problème d'optimisation. Dans la présente situation l'existence des marches aléatoires G_k crée une situation d'hétéroscédasticité, via les j – i, et introduit des pondérations non uniformes dans le problème d'optimisation. Quelle est l'interprétation financière de ces coefficients ? Ce point est important car, comme on le verra par la suite, les problèmes étudiés dans cette thèse produiront parfois des expressions mathématiques complexes. L'absence d'interprétation des différentes grandeurs intervenant dans les équations pourrait compliquer la manipulation des formules, car celles-ci ne seraient alors plus que des objets mathématiques vides de sens. La lecture financière systématique des différents

[8] $\sum$ est une matrice diagonale dont la dimension est égale à la taille de l'échantillon d'estimation. En toute rigueur on a $\hat{\Sigma}_{k,k} = 2\ \hat{\sigma}_N^2 + \hat{\sigma}_G^2\ (t_j(k) - t_i(k))$ et pas $\sum_{k,k} = 2\ \sigma_N^2 + \sigma_G^2\ (t_j(k) - t_i(k))$

éléments, au contraire, permettra bien souvent de s'orienter dans le maquis des formules.

Une vente répétée se compose d'un achat et d'une revente. Pour chaque transaction les imperfections du marché immobilier, représentées par $N_{k,t}$, écartent le prix réel de la valeur indicielle. Les données peuvent donc être comprises comme une version bruitée des niveaux de l'indice, et la variance σ_N^2 comme une mesure de la précision (ou plutôt de l'imprécision) de l'information fournie. Le bien étant négocié deux fois, les imperfections du marché s'appliqueront aussi doublement ; d'où le 2 devant σ_N^2 dans l'expression des pondérations.
Le bruit présente aussi une composante variable temporellement[9], décrite par les G_k. Au niveau local, par exemple une commune, il peut exister une tendance particulière à un quartier[10] qui n'est pas pertinente pour le niveau agrégé de l'indice. Dans cette zone, au fur et à mesure du temps, les prix deviennent de moins en moins informatifs pour l'indice, en raison de leurs évolutions idiosyncratiques. Le terme $\sigma_G^2 (j - i)$ est une mesure de la perte de précision progressive due à ce phénomène.

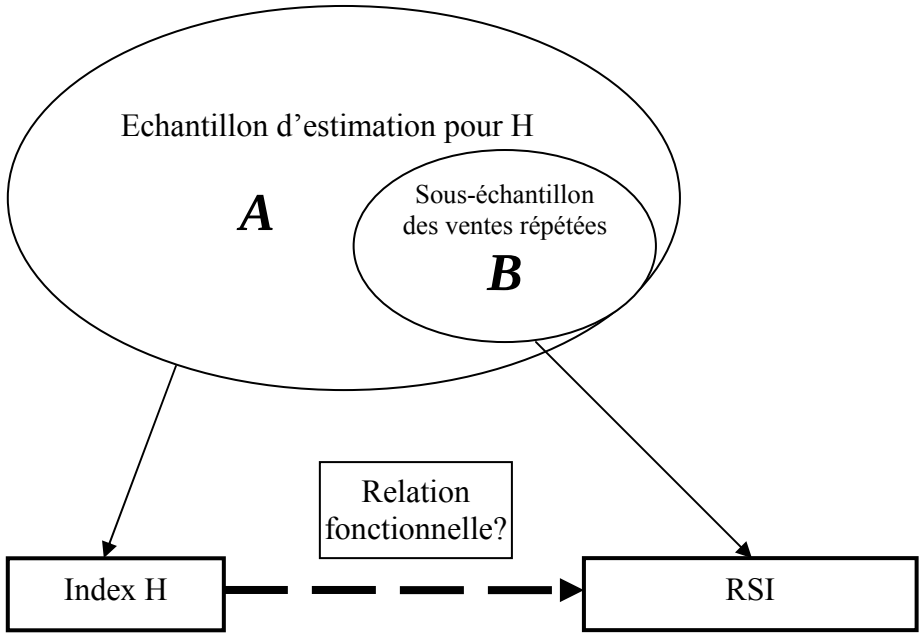
Ainsi, l'expression $2\sigma_N^2 + \sigma_G^2 (j - i)$ quantifie pour chaque vente répétée la quantité de bruit associée. Les coefficients intervenant dans le programme d'optimisation sont les inverses de ces mesures de bruit, ils évoluent donc en sens opposé. Si le bruit grandit ils diminuent et, inversement, si le bruit s'affaiblit ils deviennent plus importants. A partir de cette remarque anodine, l'interprétation des coefficients de pondération $1/(2\sigma_N^2 + \sigma_G^2 (j - i))$ devient maintenant claire. L'inverse de la variance du résidu ε_k est en fait une mesure de l'information apportée par chaque couple de ventes répétées. Cette notion de quantité d'information apparaîtra comme étant la notion centrale lors de la reformulation de l'indice de ventes répétées.

[9] En un sens l'introduction de $G_{k,t}$ présente donc une certaine dimension hédonique car on cherche alors à capturer indirectement les effets du régresseur "temps".
[10] Une décision de politique urbaine peut par exemple avoir des conséquences importantes pour une zone restreinte si des travaux d'aménagements sont décidés (construction d'un tramway, rénovation

2.3. Les intuitions : Pourquoi et où chercher une relation fonctionnelle ?

A partir d'un échantillon de transactions A on peut construire un indice de prix H (médian ou hédonique par exemple). Dans cette base de données, pour certains biens deux prix consécutifs sont connus (un prix d'achat et un prix de revente). Ce sous-échantillon B permet de construire un deuxième indice avec la technique des ventes répétées (RSI). La question étudiée ici peut s'énoncer de la manière suivante : L'inclusion du sous-échantillon B dans l'échantillon A se traduit-elle par une relation fonctionnelle entre H et R ? (figure 1). Si l'on regarde cette situation d'un point de vue informationnel, tous les renseignements de B se retrouvent dans A ; en d'autres termes le premier ensemble est « expliqué ou déterminé » par l'information enchâssée dans le second. Si ce lien est maintenu au niveau des indices, cela signifierait que le RSI peut se « déduire de H » par une formule du type RSI = F(H).

Figure 1: D'une inclusion à une relation fonctionnelle



d'un quartier, création de zones piétonnes, …). Une tendance est alors créée mais elle ne sera pertinente qu'au niveau local du quartier.

Cette intuition est à rapprocher d'un résultat bien connu en théorie de la mesure et qui est parfois utilisé pour étudier des espérances conditionnelles :

Toute fonction borélienne g : $(\Omega, \mathcal{A}) \rightarrow (R, \mathcal{B}_R)$ engendre une sous-tribu de $\mathcal{A}$.

Si : une fonction borélienne f : $(\Omega, \mathcal{A}) \rightarrow (R, \mathcal{B}_R)$ est aussi g-mesurable

alors : il existe une troisième fonction h borélienne vérifiant $f = h \circ g$.

La tribu engendrée par g s'interprète comme l'information fournie par le signal g. Affirmer que f est g-mesurable revient à considérer que la g-information est incluse dans la f-information ou plus simplement que l'observation fournie par g est plus précise que celle fournie par f. La proposition affirme alors que dans cette situation, la fonction f s'exprime comme une fonction déterministe de la fonction g, via une troisième fonction h. Pour être laconique : connaître g c'est connaître aussi f. La notion de tribu n'a pas été définie précisément sur les différents échantillons d'estimation mais si l'on remplace g par H, f par R et h par F on peut s'apercevoir qu'il s'agit bien de la même idée.

Si cette réflexion plaide pour l'existence d'un lien entre les deux indices elle n'indique toutefois pas comment la trouver. Pour cela, on peut aborder le problème sous un autre angle. Comme il a déjà été mentionné dans l'introduction, les valeurs de H sont exclusivement centrées sur une date t, tandis que le RSI utilise des intervalles de longueurs variables puisque l'on travaille avec des rendements. Pour un bien particulier, le taux de rendement est une fonction du prix d'achat et du prix de revente (figure 2). Etant donné que $p_{k,i}$ [11] est constitutif de la valeur moyenne h_i d'un indice de prix à cette même date et que, $p_{k,j}$ contribue de même à la valeur h_j,

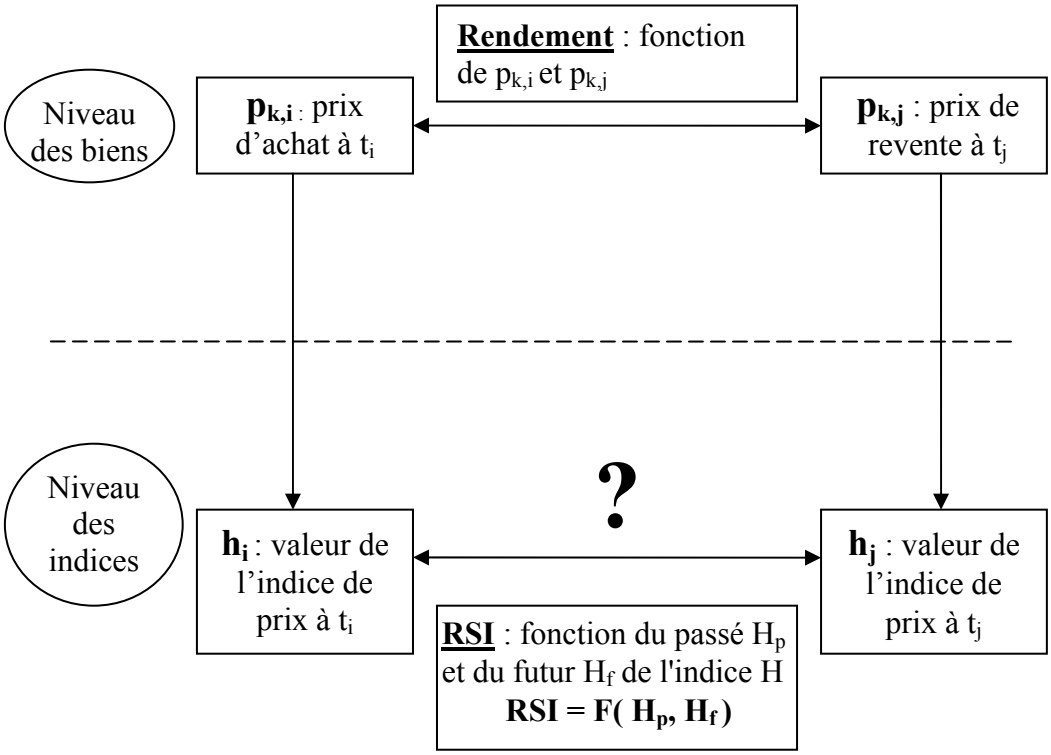
[11] $p_{k,i}$ est le prix du k^{eme} bien à t_i

l'indice de ventes répétées pourrait alors aussi être une fonction des valeurs passées et futures de l'indice de prix[12].

RSI = F(passé de H, futur de H) ?

Cette idée amènera en fait à définir la notion de moyenne des taux moyens P dans le paragraphe 3.3 qui fera le lien entre R et H.

Figure 2 : Le RSI est-il une fonction des valeurs passées et futures de H ?

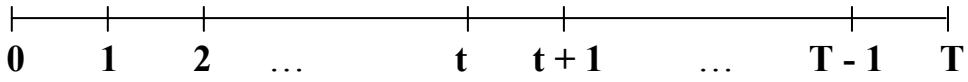


[12] En des termes plus mathématiques il s'agit de savoir si la formule du rendement passe au quotient pour les indices.

2.4. Les premières notations

2.4.1. La structure temporelle

Nous supposerons que le temps est discret et divisé en T sous-intervalles pour des dates allant de 0 à T ; la dernière représentant le présent.



On supposera également que les ventes ne se produisent qu'à ces instants et pas entre deux dates. Le pas temporel peut être le mois, le trimestre ou le semestre en fonction de la qualité des données.

2.4.2. La distribution des ventes répétées

Tableau 1: Distribution de l'échantillon des ventes répétées

Date d'achat \ Date de revente	0	1	2	3	…	t	t + 1	…	T – 2	T – 1	T
0		$n_{0,1}$	$n_{0,2}$	$n_{0,3}$		$n_{0,t}$	$n_{0,t+1}$		$n_{0,T-2}$	$n_{0,T-1}$	$n_{0,T}$
1			$n_{1,2}$	$n_{1,3}$		$n_{1,t}$	$n_{1,t+1}$		$n_{1,T-2}$	$n_{1,T-1}$	$n_{1,T}$
2				$n_{2,3}$		$n_{2,t}$	$n_{2,t+1}$		$n_{2,T-2}$	$n_{2,T-1}$	$n_{2,T}$
3						$n_{3,t}$	$n_{3,t+1}$		$n_{3,T-2}$	$n_{3,T-1}$	$n_{3,T}$
⋮											
t							$n_{t,t+1}$		$n_{t,T-2}$	$n_{t,T-1}$	$n_{t,T}$
t + 1									$n_{t+1,T-2}$	$n_{t+1,T-1}$	$n_{t+1,T}$
⋮											
T – 2										$n_{T-2,T-1}$	$n_{T-2,T}$
T – 1											$n_{T-1,T}$
T											

$n_{i,j}$: nombre de ventes répétées dans l'échantillon pour lesquels la date d'achat est à "i" et la date de revente à "j"

Chaque vente répétée génère un couple de dates (t_i ; t_j) avec $0 \leq t_i < t_j \leq T$; le nombre de possibilités pour la période de détention est de T*(T+1). Si une propriété est vendue plus de deux fois, le nombre de paires sera égal au nombre de périodes de détention. Une maison vendue par exemple quatre fois fournira trois couples. Le nombre de ventes répétées pour un couple (t_i ; t_j) sera noté $n_{i,j}$ et N désignera le nombre total de couples dans l'échantillon : $N = \sum_{i<j} n_{ij}$.

2.4.3. Les notations indicielles

Les valeurs de tout indice de prix seront notées par $H = \{ h_0, h_1, \ldots, h_T \}$; pour simplifier on prendra $h_0 = 1$ ou $h_0 = 100$. Le RSI se focalisant sur les rendements plutôt que sur les prix, on ne le décrira donc pas en termes de niveaux absolus mais en termes de taux de croissance. r_t désignera ainsi le taux de rendement de cet indice sur la période unitaire [t,t+1] comme indiqué précédemment.

En notant $H = \{ h_0, h_1, \ldots, h_T \}$ et $R = \{ r_0, r_1, \ldots, r_{T-1} \}$ l'éventuelle relation « déterministe » entre ces deux indices s'écrira donc :

$$R = \{r_0, r_1, \ldots, r_{T-1}\} = \{ \mathbf{f_0}(h_0, h_1, \ldots, h_T), \mathbf{f_1}(h_0, h_1, \ldots, h_T), \ldots, \mathbf{f_{T-1}}(h_0, h_1, \ldots, h_T) \}$$
$$= \{ \mathbf{f_0}(H), \mathbf{f_1}(H), \ldots, \mathbf{f_{T-1}}(H)\} = \mathbf{F}(H)$$

avec une fonction $\mathbf{F} = (f_0, f_1, \ldots, f_{T-1})$ à préciser.

3. Exploration du problème dans un environnement simplifié

Ce paragraphe résout le problème du lien fonctionnel entre les deux indices dans un contexte élémentaire. On verra apparaître au cours de l'étude la structure fine du RSI qui fera l'objet d'une généralisation ultérieure dans le paragraphe 6 de ce chapitre.

3.1. Les simplifications

Afin d'établir la relation fonctionnelle R = F(H), le problème va d'abord être étudié dans un cadre simplifié. On supposera que toutes les transactions se font au niveau de l'indice de prix H ; par conséquent les coordonnées du vecteur[13] Y sont simplement de la forme : $\ln(h_j / h_i)$. Cette affirmation est bien sûr fausse au cas par cas, car cela revient à ignorer les imperfections du marché. Elle peut cependant être considérée comme vraie globalement car la fonction de l'indice H est de rendre compte par un chiffre moyen d'une situation hétérogène[14].

Si l'on impose que toutes les transactions se font au niveau de l'indice H, pour être cohérent il faut alors aussi considérer que les bruits blancs $N_{k,t}$ sont inexistants ; ou en d'autres termes que la volatilité σ_N est nulle. La relation (2) devient alors :

$$\ln (h_j / h_i) = \ln(\text{Indice}_j) - \ln(\text{Indice}_i) + G_{k,j} - G_{k,i} \qquad (8)$$

[13] On rappelle que le vecteur Y est le vecteur des rendements réalisés dans l'échantillon d'estimation, cf. paragraphe 2.1

[14] Cette affirmation peut être considérée comme discutable s'il y a un divergence significative entre les tendances moyennes des deux échantillons de la Figure 1, A et B. L'indice H est par nature adapté à son échantillon d'estimation A mais il peut être imparfait pour décrire B si, par exemple, il existe un biais de sélectivité dans cet ensemble B.

Des deux sources de bruits, G et N, une seule est conservée. Cette modélisation correspond à un monde dans lequel toutes les transactions se font à leur "juste" prix mais où les tendances spécifiques à chaque bien existent toujours. Cette situation est contradictoire car si toutes les transactions se font au niveau de l'indice H, non seulement les bruits blancs sont nuls, mais il devrait en être de même pour les marches aléatoires G ; on devrait donc aussi avoir $\sigma_G = 0$. Cette écriture artificielle du problème n'est nullement définitive car il ne s'agit ici que d'un procédé exploratoire. Dans le paragraphe 6 de ce chapitre un résultat général reposant sur une modélisation non contradictoire sera énoncé mais, à ce stade de la démarche, pour comprendre la mécanique des formules et l'articulation des concepts, on raisonnera sur une situation irréaliste financièrement mais simple mathématiquement. On peut toutefois conserver un minimum d'interprétation économique pour cette modélisation en considérant que poser le problème en ces termes revient à affirmer que les transactions se font bien au niveau de l'indice H, mais que plus la détention est longue, moins cette affirmation est sûre (car la variance du terme $G_{k,j} - G_{k,i}$, $\sigma_G^2 (j-i)$, augmente avec le temps).

3.2. Echantillon d'estimation et contribution informationnelle

Si l'on essaye de retrouver intuitivement le taux r_t du RSI pour l'intervalle [t,t+1], les couples pertinents sont en premier lieu, ceux dont la date d'achat est t et la date de revente t + 1. En utilisant les notations introduites ci-dessus, on obtient donc $n_{t,t+1}$ paires, correspondant à une seule cellule du tableau 1. Mais choisir seulement ces couples pour estimer r_t est un mauvaise option, car on gaspille alors une part importante de l'information contenue dans l'échantillon. Toutes les ventes répétées dont la période de détention est par exemple [t-1, t+1] ne seront pas utilisées, alors qu'elles contiennent de l'information sur l'évolution des prix de l'immobilier entre t et t + 1. On pourrait de plus être confronté à un problème de rareté des données pour

certains intervalles élémentaires [t,t+1], car il est assez exceptionnel qu'un bien soit détenu seulement sur une période.

La bonne solution consiste en fait à choisir toutes les paires $(t_i ; t_j)$ telles que $t_i \leq t$ et $t_j \geq t + 1$ (achat avant t, revente après t+1). Le tableau 2 représente les sous-échantillons correspondant à chacun de ces deux choix et il est évident que le second (traits pleins) est préférable au premier (pointillés) ; la quantité de données étant beaucoup plus importante. On notera Spl^t le sous échantillon des ventes répétées pertinentes pour l'intervalle [t , t + 1] correspondant au rectangle en traits pleins.

Tableau 2 : échantillon d'estimation pour r_t

Date d'achat \ Date de revente	0	…	t	t + 1	…	T
0				$n_{0,t+1}$		$n_{0,T}$
⋮						
t				$n_{t,t+1}$		$n_{t,T}$
t + 1						
⋮						
T						

Traits pleins : achat avant t et revente à partir de t+1
Pointillés : achat à t, revente à t+1

Si le problème des données rares peut se gérer avec un meilleur choix de l'échantillon d'estimation, toutes les paires de celui-ci n'auront pas le même niveau hiérarchique. On verra en effet par la suite que, plus une cellule sera éloignée de la diagonale, moins sa contribution informationnelle sera forte.[15].

[15] La mesure d'information pour une cellule, $L_{i,j} = n_{i,j} / (j – i)$, définie ci-après sera d'autant plus faible que l'on s'éloignera de la diagonale, ou en d'autres termes que la période de détention (j – i) augmentera.

3.3. La moyenne des taux moyens

Le concept étudié dans cette partie, ρ_t, représente le taux moyen réalisé par les investisseurs qui détenaient de l'immobilier pendant l'intervalle [t,t+1]. On peut pressentir que cette grandeur doit probablement être très liée aux valeurs passées et futures de H (cf. figure 2). De plus, elle devrait avoir un certain lien avec r_t , le taux de croissance du RSI sur ce même intervalle.

Pour définir ce concept on sélectionne toutes les ventes répétées (t_i ; t_j)[16] pertinentes pour l'estimation de r_t, c'est-à-dire celles de Spl^t. On note n^t le nombre total de ces données :

$$n^t = \sum_{i \le t < j} n_{i,j} \qquad (t = 0, 1, \ldots, T-1) \qquad (9)$$

Pour chaque couple (t_i , t_j), on a supposé que les transactions se font au niveau de H. Les rendements sont donc identiques pour tous les biens d'une même cellule, à savoir h_j / h_i. Si l'on cherche le taux de rentabilité continu, moyen, réalisé sur cette période de détention, il faut résoudre :

$$e^{r_{i,j}\,(j-i)} = h_j / h_i \quad \Leftrightarrow \quad r_{i,j} = (\ln h_j - \ln h_i) / (j - i) \qquad (10)$$

L'information fournie pas ce taux constant $r_{i,j}$ ne concerne pas seulement r_t mais aussi tous les autre taux entre r_i et r_j ; on ne peut donc pas l'assimiler directement à r_t. Toutefois à partir de ces valeurs on peut définir le concept intermédiaire de moyenne des taux moyens, noté ρ_t, en calculant la moyenne des $r_{i,j}$ pour les cellules constituant le rectangle du tableau 2 :

[16] Dans un souci de simplicité (t_i ; t_j) sera aussi parfois noté (i , j) ; j – i sera alors la longueur de la période de détention.

$$\rho_t = \sum_{i \le t < j} (n_{i,j} / n^t) \; r_{i,j} \qquad (11)$$

ρ_t est une moyenne arithmétique équipondérée représentant la profitabilité moyenne du foncier pour la population des investisseurs qui détenaient de l'immobilier entre t et t + 1. Il se réécrit :

$$\begin{aligned} \rho_t &= (1 / n^t) \sum_{i \le t < j} n_{i,j} \; \ln [(h_j / h_i)^{1 / (j - i)}] \\ &= (1 / n^t) \sum_{i \le t < j} \ln [(h_j / h_i)^{n_{i,j} / (j - i)}] \\ &= (1 / n^t) \; \ln [\prod_{i \le t < j} (h_j / h_i)^{n_{i,j} / (j - i)}] \end{aligned}$$

A ce niveau l'expression formelle de ρ_t est encore difficile à manipuler et l'interprétation ne semble pas évidente. Pour apporter de la lisibilité à cette formule on peut remarquer que, puisque tous les achats sont avant t et toute les reventes après t, ρ_t devrait être une sorte de division entre les valeurs passées de H et ses valeurs futures. Concrètement, le double produit des quotients h_j / h_i peut s'exprimer comme un quotient de deux produits ; le numérateur concernant uniquement le futur de H et le dénominateur le passé. Afin d'alléger l'expression, on notera par $L_{i,j}$ les quantités $n_{i,j} / (j - i)$; l'interprétation financière de cette grandeur sera développée par la suite.

On a donc : $$\rho_t = (1 / n^t) \; \ln \Big[\prod_{i \le t < j} (h_j^{L_{i,j}} / h_i^{L_{i,j}}) \Big]$$

Une double indexation n'est jamais très intuitive, mais si l'on remarque qu'il y a autant de couples (i , j) que de cellules à l'intérieur du rectangle du tableau 2, les choses deviennent plus simples. Pour un j fixé, l'exposant de h_j est simplement $L_{0,j} + L_{1,j} + \ldots + L_{t,j}$ et il correspond en fait à toutes les possibilités d'acheter avant t

et de revendre exactement à $t = t_j$. De même pour un i fixé, l'exposant de h_i est $L_{i,t+1} + L_{i,t+2} + \ldots + L_{i,T}$, il correspond à toutes les possibilités d'acheter exactement à $t = t_i$ et de revendre à $t + 1$ ou après.

En utilisant les notations[17]: $B_i^t = L_{i,t+1} + L_{i,t+2} + \ldots + L_{i,T}$ et $S_j^t = L_{0,j} + L_{1,j} + \ldots + L_{t,j}$

on obtient : $$\rho_t = (1 / n^t) \ln \left[\left(\prod_{j > t} h_j^{S_j^t}\right) / \left(\prod_{i \le t} h_i^{B_i^t}\right) \right]$$

Ces deux produits peuvent s'interpréter comme des moyennes géométriques pondérées des valeurs de H. On rappelle ici que la moyenne géométrique[18] des $\{x_1 , x_2 , \ldots , x_n \}$ avec les poids $(\alpha_1 , \alpha_2 , \ldots , \alpha_n)$ est le nombre G vérifiant :

$$G^{\alpha} = x_1^{\alpha_1} x_2^{\alpha_2} \ldots x_n^{\alpha_n}$$ où $\alpha = \alpha_1 + \alpha_2 + \ldots + \alpha_n$ représente le poids total

Introduisons donc la moyenne géométrique $H_p(t)$ des $\{h_0, h_1, \ldots , h_t\}$ pondérés par les $\{B_0^t, B_1^t, \ldots ,B_t^t\}$ et la moyenne $H_f(t)$ des $\{h_{t+1}, \ldots ,h_T\}$ avec les poids $\{S_{t+1}^t,\ldots,S_T^t\}$.

La pondération totale est : $$B^t = \sum_{i = 0,\ldots,t} B_i^t \qquad S^t = \sum_{j = t+1,\ldots,T} S_j^t$$

Mais comme on a : $$B^t = \sum_{i = 0,\ldots,t} B_i^t = \sum_{i=0,\ldots,t} \sum_{j = t+1,\ldots,T} L_{i,j}$$

$$= \sum_{j=t+1,\ldots,T} \sum_{i=0,\ldots,t} L_{i,j} = \sum_{j=t+1,\ldots,T} S_j^t = S^t$$

les poids totaux pour les deux moyennes géométriques sont en fait les mêmes. On utilisera donc une notation commune : $$B^t = S^t = I^t \qquad (12)$$

[17] B_i pour "buy at t_i" et S_j pour "sell at t_j"

Il vient alors : $$[H_f(t)]^{I^t} = \prod_{j>t} h_j^{S_j^t} \qquad [H_p(t)]^{I^t} = \prod_{i \le t} h_i^{B_i^t} \qquad (13)$$

et on peut donc écrire : $$\rho_t = (1 / n^t)\ \ln (H_f(t)^{I^t} / H_p(t)^{I^t})$$

$$\rho_t = (I^t / n^t) * \ln \left[H_f(t) / H_p(t) \right] \qquad (14)$$

L'objectif annoncé d'établir une relation du type RSI = F(H_p, H_f) n'est pas encore atteint pour le vecteur R, mais il est maintenant réalisé pour le vecteur des taux moyens P = (ρ_0, …, ρ_{T-1}), via la formule 14

3.4. Les interprétations informationnelles des différentes grandeurs introduites

Tous ces concepts introduits ci-dessus ($L_{i,j}$, B_i^t, S_j^t, I^t, H_f et H_p) ne sont pas purement formels, ils peuvent être interprétés en termes d'informationnels et financiers.

3.4.1. $L_{ij} = n_{i,j} / (j - i)$: contribution informationnelle d'une cellule (t_i , t_j)

Pour une donnée k, on a pu voir dans le paragraphe 2.1 que la variance du résidu ε_k s'écrivait $2\sigma_N^2 + \sigma_G^2 (j - i)$ et que cette quantité pouvait s'interpréter comme une mesure de bruit ; son inverse $1 / (2\sigma_N^2 + \sigma_G^2(j - i))$ étant alors une mesure d'information. En intégrant la simplification introduite dans le paragraphe 3.1 ($\sigma_N = 0$), cette quantité devient $1/(\sigma_G^2(j - i))$. Si l'on considère de plus que

[18] Pour plus de détails sur les concepts de moyennes on se reportera à l'annexe 13.

l'information est définie à une constante multiplicative près[19], celle fournie par une vente répétée réalisée entre "i" et "j" est alors simplement égale à 1 / (j – i). Les $n_{i,j}$ données associées à ces deux dates délivreront donc une quantité d'information totale de $n_{i,j}$ x 1/(j – i) = $n_{i,j}$ /(j – i) = $L_{i,j}$. $L_{i,j}$ s'interprétera donc comme la contribution informationnelle pour la classe (i,j) des ventes répétées.

3.4.2. B_i^t, S_j^t, B^t, S^t, I^t : indicateurs de la distribution informationnelle

$B_i^t = L_{i,t+1} + L_{i,t+2} + \ldots + L_{i,T}$ concerne les transactions dont la date d'achat est t_i et la date de revente postérieure à t. Cette somme ne porte pas sur les nombres de transactions réelles ($n_{i,j}$) mais sur les équivalents informationnels ($L_{i,j}$). B_i^t peut donc être perçue comme la contribution informationnelle des ventes répétées réalisées entre t_i et les $t_j > t$. De même $S_j^t = L_{0,j} + L_{1,j} + \ldots + L_{t,j}$ quantifie l'information fournie par les couples dont la date d'achat est avant t et la revente exactement à t_j.

Dans le même esprit, $B^t = B_0^t + \ldots + B_t^t$ donne la contribution informationnelle de toutes les transactions effectuées avec $t_i \leq t$ et $t_j > t$; il s'agit en fait simplement des cases à l'intérieur du rectangle du tableau 2. Quant à $S^t = S_{t+1}^t + \ldots + S_T^t$, bien que l'expression semble différente à première vue, il s'agit de la même quantité. Le calcul pour B^t se fait d'abord par sommation horizontale puis verticale, tandis que pour S^t l'ordre est inversé. La formule (12), $B^t = S^t = I^t$, ne fait alors qu'exprimer cette simple idée.

D'une manière générale I^t est donc la mesure de la quantité d'information pertinente pour l'intervalle [t ; t + 1] ; elle s'obtient à partir des couples dont la période de détention englobe cet intervalle.

[19] Cette hypothèse ne semble pas déraisonnable car il n'existe pas vraiment d'unité de mesure de l'information. Ce qui importe n'est pas tant le niveau absolu d'information pour une donnée que son niveau relatif par rapport aux autres données.

3.4.3. $H_f(t)$, $H_p(t)$: moyennes des prix immobiliers passés et futurs, pondérés par l'activité informationnelle du marché

Les formules (13) définissent deux moyennes pour H : $H_p(t)$ et $H_f(t)$. $H_p(t)$ concerne les valeurs passées[20] et les poids de chacun des h_i correspondent aux quantités d'information fournies par les ventes répétées pour lesquelles l'achat a été réalisé aux dates t_i correspondantes. Ces exposants sont constitués à partir de deux éléments, la période de détention j – i et le nombre réel des transactions $n_{i,j}$, qui agissent dans des directions opposées. Pour un h_i particulier, si la période de détention moyenne s'accroît, comme j – i est au dénominateur dans $L_{i,j}$, l'exposant B_i^t diminuera réduisant ainsi l'influence de h_i dans $H_p(t)$. Par contre si $n_{i,j}$ augmente, la somme B_i^t évoluera dans le même sens, renforçant la contribution de h_i à la moyenne. Par conséquent selon les niveaux d'activité[21] du marché, déterminés par les choix des dates d'achat et de revente par les propriétaires, les niveaux des moyennes $H_f(t)$ et $H_p(t)$ varieront.

Pour le sous-échantillon des transactions pertinentes pour l'estimation du taux de rendement de l'indice sur [t, t +1], $H_p(t)$ peut s'interpréter comme le prix fictif d'achat et $H_f(t)$ comme le prix fictif de revente. Une structure géométrique des moyennes n'est pas vraiment surprenante ; il a en effet déjà été pointé par Shiller (1991) ou Goetzmann (1992) que les indices de ventes répétées produisaient des estimateurs géométriques.

3.4.4. ρ_t : taux moyen réalisé

La première partie de la formule (14) mesure la qualité informationnelle de l'échantillon Spl^t en divisant la quantité d'information fournie, I^t, par le nombre de couples n^t. Globalement, plus la période de détention est longue, moins l'échantillon

[20] $H_f(t)$ les valeurs futures de l'indice.

est informatif. Mais, le ratio I^t / n^t a également une seconde interprétation. Le nombre $1 / (j – i)$ étant l'inverse de la période de détention, on peut l'assimiler à une fréquence[22], ou à un taux de rotation du parc immobilier. I^t est alors défini comme la somme des fréquences pour tous les couples de Spl^t. Or, comme le nombre de ces données est égal à n^t, le ratio $F^t = I^t / n^t$ n'est en fait rien d'autre qu'une moyenne arithmétique des fréquences de ce sous-échantillon. A partir de ce F^t on peut ensuite définir la période moyenne de détention par :

$$\tau^t = (F^t)^{-1} = (I^t / n^t)^{-1} = n^t / I^t$$

Il faut toutefois remarquer que τ^t n'est pas une simple moyenne arithmétique des durées de détention, il s'agit en fait d'une moyenne harmonique. On rappelle que la moyenne harmonique[23] de $\{x_1 , x_2 , \dots , x_n \}$ pondérée par les $(\alpha_1 , \alpha_2 , \dots , \alpha_n)$ est le nombre F vérifiant:

$$\alpha / F = \alpha_1 / x_1 + \dots + \alpha_n / x_n \qquad \text{avec } \alpha = \sum_{i=1,\dots, n} \alpha_i$$

Comme on a $n^t / \tau^t = I^t = \sum_{i \le t < j} n_{i,j} /(j – i)$, avec $n^t = \sum_{i \le t < j} n_{i,j}$, τ^t est donc bien la moyenne harmonique des périodes de détention.

Avec ces idées à l'esprit, la formule (14) devient maintenant limpide puisque on peut la réécrire :

$$\rho_t = (1/ \tau^t) * (\ln H_f(t) – \ln H_p(t)) \qquad (14')$$

---

[21] Activité informationnelle puisque l'on travaille avec les $L_{i,j}$, via les B_i^t

[22] En physique l'inverse d'une durée est une fréquence, mesurée en Hertz.

[23] Cf. annexe 13.

Il s'agit en fait d'une version agrégée de l'expression $r_{i,j} = (\ln h_j - \ln h_i) / (j-i)$ donnant le taux moyen réalisé pour un bien particulier négocié aux niveaux de l'indice. (j – i) est remplacé par la moyenne harmonique des périodes de détention, h_i par le prix moyen d'achat et h_j par le prix moyen de revente. Une telle relation pourrait sembler basique à première vue mais il n'en est rien car les pondérations ont été précisées explicitement dans les différentes formules et interprétées en termes de niveau d'activité informationnelle du marché (côté vente avec les S_j^t et côté achat avec les B_i^t). Cette écriture n'est donc absolument pas triviale.

Tous les concepts définis jusqu'ici sont associés à l'intervalle [t,t+1] et au rectangle du tableau 2. Ils peuvent être généralisés sans grande difficulté au cas de l'échantillon considéré dans son ensemble (cf. annexe 1).

3.5. La solution du problème de minimisation

$P = (\rho_0, \ldots, \rho_{T-1})$ est une fonction de H_p et H_f , et donc de l'indice de prix H. Pour obtenir une relation entre R et H il suffit maintenant de trouver un lien théorique entre R et P. Pour cela, on reprend la définition du vecteur R comme solution du problème de minimisation des moindres carrés. Comme on va s'en apercevoir, P apparaît naturellement dans la résolution de ce dernier.

Dans le cas simplifié, le problème d'optimisation s'écrit :

$$\text{Min}_R \left[\sum_{i<j} \{n_{i,j} / \sigma_G^2 (j - i) \} * \{\ln(h_j / h_i) - (r_i + \ldots + r_{j-1})\}^2\right] \qquad (15)$$

où les inconnues sont les valeurs $r_0, \ldots, r_{T-1}$ réunies dans le vecteur R

Comme la constante σ_G^2 peut se mettre en facteur, il devient simplement :

$$\text{Min}_R \left[\sum_{i<j} L_{i,j} * (\ln(h_j/h_i) - (r_i + \ldots r_{j-1}))^2\right] \qquad (15')$$

On peut remarquer que dans cette situation la solution est évidente. Le minimum de la fonction objectif est zéro, et il est atteint en choisissant $r_t = \ln (h_{t+1} / h_t)$ ou, en d'autres termes, Ind = H (on rappelle que Ind désigne le vecteur des niveaux de l'indice RSI). Malgré tout, même si la solution du problème est connue dès le départ, on considérera volontairement que ce n'est pas le cas. L'arrière-pensée de cette démarche consiste en fait à trouver une méthode de résolution susceptible de se généraliser car le véritable problème, énoncé en toute généralité dans le paragraphe 6 et résolu dans les annexes 6 à 10, est complexe et sa solution n'est absolument pas évidente. On profitera donc de ce cadre allégé pour comprendre les relations entre les différentes briques élémentaires constituant la solution, en espérant pouvoir les prolonger à l'environnement complexe.

En développant les carrés et en conservant uniquement les termes non constants, le problème s'écrit :

$$\text{Min}_{\{r_0, \ldots, r_{T-1}\}} \Phi(R) \qquad (15'')$$

avec: $\Phi(R) = \sum_{i<j} L_{i,j} [(r_i + \ldots r_{j-1})^2 - 2 \ln(h_j/h_i) (r_i + \ldots r_{j-1})]$

La fonction Φ est continûment différentiable deux fois et strictement convexe, le problème peut donc être résolu avec les seules conditions du premier ordre en annulant simplement les dérivées. On calcule donc pour chaque t = 0, …, T – 1 les dérivées par rapport à r_t de $\Phi(R)$. On utilisera la notation $\check{r}_t$ pour designer la somme de r_i à r_j, r_t excepté[24]. Comme r_t est présent dans la contribution d'un couple (i, j) si et seulement si $i \leq t < j$ on a :

$$\partial\Phi(R) / \partial r_t = \sum_{i \leq t < j} L_{i,j} [2 r_t + 2 \check{r}_t - 2 \ln(h_j/h_i)]$$

$$\partial\Phi(R) / \partial r_t = \sum_{i \leq t < j} 2L_{i,j} [(r_i + \ldots + r_{j-1}) - \ln(h_j/h_i)]$$

[24] $r_i + \ldots r_{j-1} = \check{r}_t + r_t$ d'où $(r_i + \ldots r_{j-1})^2 = r_t^2 + \check{r}_t^2 + 2 \check{r}_t r_t$ qui peut se dériver facilement.

$$\partial\Phi(R) / \partial r_t = 2 \sum_{i \le t < j} L_{i,j} (r_i + \ldots + r_{j-1}) - 2 \sum_{i \le t < j} L_{i,j} \ln(h_j/h_i)$$

$$\partial\Phi(R) / \partial r_t = 2 \sum_{i \le t < j} L_{i,j} (r_i + \ldots + r_{j-1}) - 2\, n^t \rho_t \qquad (16)$$

Le concept de taux moyen réalisé par les investisseurs introduit précédemment (ρ_t) apparaît donc dans le problème de minimisation. Pour la première sommation la signification est moins évidente mais, en réorganisant légèrement les calculs, on peut réussir à l'interpréter à l'aide des différents concepts définis ci-dessus (cf. annexe 2 pour les détails du calcul).

On démontre ainsi que pour $t' \le t$, le nombre total de $r_{t'}$ dans le somme est :

$$B_0^t + B_1^t + \ldots + B_{t'}^t$$

et que pour $t' > t$, le nombre total de $r_{t'}$ est :

$$S_T^t + S_{T-1}^t + \ldots + S_{t'+1}^t$$

Ces deux quantités amènent à définir un nouveau concept, généralisation naturelle de I^t. Pour $t' \le t$, les ventes répétées apportant de l'information sur l'intervalle temporel $[t', t+1]$ sont celles dont la date d'achat est à t' ou avant, et la date de revente à $t+1$ ou après, comme représenté dans le tableau 3. La quantité d'information pertinente pour $[t', t+1]$, notée $I^{[t', t+1]}$, peut se calculer indifféremment par $B_0^t + B_1^t + \ldots + B_{t'}^t$ (côté achat), ou par $S_T^{t'} + S_{T-1}^{t'} + \ldots + S_{t+1}^{t'}$ (côté vente), et on a évidemment :

$$I^t = I^{[t, t+1]} = B_0^t + B_1^t + \ldots + B_t^t = S_T^t + S_{T-1}^t + \ldots + S_{t+1}^t$$

$I^{[t', t+1]}$ étant un concept informationnel on peut définir immédiatement son équivalent réel, noté $n^{[t', t+1]}$, par $b_0^t + b_1^t + \ldots + b_{t'}^t$ (ou de manière équivalente par le côté vente avec $s_T^{t'} + s_{T-1}^{t'} + \ldots + s_{t+1}^{t'}$).

Tableau 3 : Données pertinentes et quantité d'information associées à l'intervalle [t',t+1]

	0	…	t'	…	t	t+1		T	**Somme**
0			$L_{0,\,t'}$		$L_{0,\,t}$	$L_{0,\,t+1}$		$L_{0,\,T}$	B_0^t
⋮									⋮
t'					$L_{t',\,t}$	$L_{t',\,t+1}$		$L_{t',\,T}$	$B_{t'}^t$
⋮									⋮
t						$L_{t,\,t+1}$		$L_{t,\,T}$	⋮
⋮									⋮
T									⋮
					Somme	$S^{t'}_{t+1}$	…	$S^{t'}_{T}$	$I^{[t',t+1]}$

$\partial\Phi(R) / \partial r_t$ s'écrit donc maintenant :

$$\partial\Phi(R) / \partial r_t = 2 \sum_{t' \le t} I^{[\,t',\,t+1]}\, r_{t'} + 2 \sum_{t' > t} I^{[\,t,\,t'+1]}\, r_{t'} - 2\, n^t\, \rho_t \qquad (16')$$

Introduisons la matrice carrée Î, de dimension T :

$$\hat{I} = \begin{pmatrix} I^{[0,1]} & I^{[0,2]} & I^{[0,3]} & & I^{[0,T]} \\ I^{[0,2]} & I^{[1,2]} & I^{[1,3]} & & I^{[1,T]} \\ I^{[0,3]} & I^{[1,3]} & I^{[2,3]} & & I^{[2,T]} \\ \vdots & & & & \\ I^{[0,T]} & I^{[1,T]} & I^{[2,T]} & & I^{[T-1,T]} \end{pmatrix}$$

Pour $p \leq q$, la (p,q) composante[25] $\hat{I}_{p,q}$ de $\hat{I}$, est $I^{[p-1,\, q]}$ et pour $p > q$, $I^{[q-1,\, p]}$. En termes concrets si l'on choisit par exemple l'intervalle temporel [2; 5], la quantité d'information pertinente associée, à savoir $I^{[2,\, 5]}$, est fournie par $\hat{I}_{3,5}$ ou par $\hat{I}_{5,3}$. Comme on peut lire directement à partir de la matrice $\hat{I}$ la quantité d'information correspondant à n'importe quel intervalle, on appellera cette matrice symétrique, la matrice d'information pertinente[26].

Si l'on définit de plus une matrice diagonale de dimension T, notée η, dont les valeurs diagonales sont $n^0, n^1, \ldots, n^{T-1}$, les T équations du système des conditions du premier ordre $\{ \partial\Phi(R) / \partial r_t = 0 \; ; \; t = 0, \ldots, T-1 \}$ se réécrivent alors simplement :

$$\hat{I}\, R = \eta\, P \qquad (17)$$

Et si l'inverse de $\hat{I}$ existe, la solution de cette équation d'inconnue R (vecteur des taux de croissance monopériodiques de l'indice) est finalement :

$$R = (\hat{I}^{-1} \eta)\; P \qquad (17')$$

Le produit matriciel $(\hat{I}^{-1} \eta)$ fournit donc le lien entre les taux R du RSI et les taux moyens réalisés $P = (\rho_0, \rho_1, \ldots, \rho_{T-1})$. La question de départ – 'Existe-t-il un lien fonctionnel entre R et H? ' – peut maintenant être résolue.

[25] On indexe la matrice par p et q avec $1 \leq p,q \leq T$

[26] Elle possède les propriétés élémentaires suivantes :

- composantes positives et symétriques : $\hat{I}_{p,q} \geq 0 \qquad \hat{I}_{p,q} = \hat{I}_{q,p}$
- Les termes décroissent en ligne et en colonne à partir des éléments diagonaux :

  $\hat{I}_{p,p} \geq \hat{I}_{p,p+1} \geq \ldots \geq \hat{I}_{p,T}$ et $\hat{I}_{p,0} \leq \hat{I}_{p,1} \leq \ldots \leq \hat{I}_{p,p}$ pour p = 0, …, T

  $\hat{I}_{p,p} \geq \hat{I}_{p+1,p} \geq \ldots \geq \hat{I}_{T,p}$ et $\hat{I}_{0,p} \leq \hat{I}_{1,p} \leq \ldots \leq \hat{I}_{p,p}$ pour p = 0, …, T

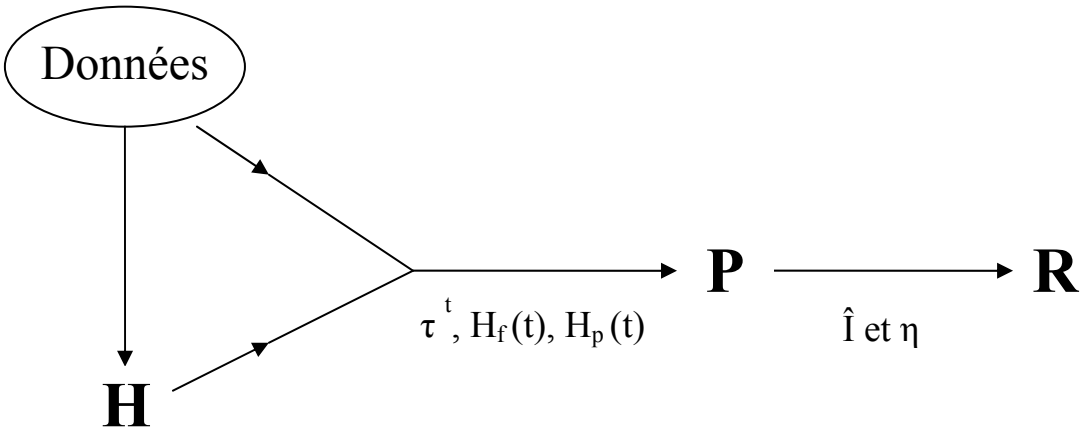
3.6. La relation fonctionnelle entre les deux indices : Bilan

Comme illustré dans la figure 3, les taux moyens réalisés par les investisseurs pour chaque période, $P = (\rho_0, \rho_1, \ldots, \rho_{T-1})$, constituent le chaînon entre l'indice de prix H et l'indice de ventes répétées R (ou Ind en niveau). Dans un premier temps on obtient P à partir des données en calculant les périodes moyennes de détention τ^t, ainsi que les moyennes passées et futures de H, $H_f(t)$ et $H_p(t)$. P représente le taux fictif réalisé si l'on achète au niveau $H_p(t)$ et que l'on revend au niveau $H_f(t)$, en ayant détenu le bien pendant une durée de τ^t (ces différentes grandeurs sont pondérées par les niveaux d'activité du marché). Puis, dans un second temps, on déduit R de P en utilisant la matrice informationnelle $\hat{I}$. Le premier élément de (17'), $\hat{I}^{-1}\eta$, dépend seulement de la distribution des dates des ventes répétées, et les valeurs de l'indice H n'y apparaissent pas directement. Pour autant, cela ne signifie pas qu'elles n'ont pas d'influence sur cette matrice, car H peut agir indirectement[27]. Mais, pour résumer, comme P est une fonction de H la relation $R = (\hat{I}^{-1}\eta) P$ implique qu'il en est de même pour R. La question de la relation fonctionnelle entre ces deux indices est donc maintenant résolue.

Cette relation a été établie, pour des raisons de simplicité d'exposition, dans un cas trivial où les deux indices coïncidaient exactement. Cependant, comme on le verra dans le paragraphe 6 et dans les annexes pour la démonstration en toute généralité, la structure fondamentale du lien que l'on observe ici sera conservée dans le cas non trivial.

[27] Si l'on se trouve dans une situation où les prix immobiliers sont très bas, on peut raisonnablement supposer que le niveau de liquidité est également faible (les propriétaires préférant attendre de meilleurs jours avant de vendre leur bien). Dans une telle configuration, les dates de reventes sont repoussées en raison des valeurs de H ; le niveau des prix affecte donc la matrice $\hat{I}$.

Figure 3 : Relation entre R et H



Au cours de cette recherche sur l'existence d'une relation déterministe entre les deux indices, la structure fine du RSI a été mise en évidence ; le concept de contenu informationnel d'une transaction est ainsi apparu comme LE concept central. Ce résultat invite à reformuler plus globalement l'indice de ventes répétées en s'appuyant sur le formalisme qui a émergé au cours de cette étude. L'indice RSI ne serait plus alors défini comme le résultat d'une simple régression, où la structure d'information serait implicite, mais comme le résultat d'un calcul informationnel explicite. Cette nouvelle approche sera présentée dans le paragraphe 6 sous une forme déductive et algorithmique, et non plus exploratoire comme dans le présent paragraphe. Elle permettra par la suite d'approfondir l'analyse des données immobilières (chapitre 2), de résoudre certains problèmes classiques pour le RSI (chapitre 3) et d'ouvrir des pistes de recherche vers ce que l'on pourrait appeler une théorie des indices informationnels (chapitre 4).

3.7. Une variation dans l'énoncé des résultats

Si l'on remplace dans (16') $n^t\rho_t$ par $I^t \ln [H_f(t) / H_p(t)]$ et si l'on divise chaque équation $\partial\Phi(R)/\partial r_t = 0$ par I^t, comme $I^{[t, t+1]} = I^t$, on obtient :

$$\sum_{t' < t} (I^{[t', t+1]} / I^{t}) \, r_{t'} + r_t + \sum_{t' > t} (I^{[t, t'+1]} / I^{t}) \, r_{t'} = \ln [H_f(t) / H_p(t)]$$

Introduisons alors J, la T*T-matrice :

$$J = \begin{pmatrix} 1 & I^{[0,2]}/I^{0} & I^{[0,3]}/I^{0} & \cdots & I^{[0,T]}/I^{0} \\ I^{[0,2]}/I^{1} & 1 & I^{[1,3]}/I^{1} & & I^{[1,T]}/I^{1} \\ I^{[0,3]}/I^{2} & I^{[1,3]}/I^{2} & 1 & & I^{[2,T]}/I^{2} \\ \vdots & & & & \\ I^{[0,T]}/I^{T-1} & I^{[1,T]}/I^{T-1} & I^{[2,T]}/I^{T-1} & & 1 \end{pmatrix}$$

et Y le vecteur colonne dont les composantes sont $\ln[H_f(t) / H_p(t)]$ $(t = 0, \ldots, T-1)$.

Comme $\ln[H_f(t) / H_p(t)] = \tau^{t} \rho_t$, Y est simplement, pour l'ensemble des propriétaires détenant de l'immobilier entre t et t + 1, le rendement moyen réalisé sur toute leur période de détention

Les T équations de (17) s'écrivent alors : $J R = Y$ (18)

Et si l'inverse de J existe : $R = J^{-1} Y$ (18')

La matrice J peut s'interpréter[28] de la façon suivante. On prend ici l'exemple de la troisième ligne qui est associée à l'intervalle [2,3], la quantité d'information correspondante est $I^2 = I^{[2,3]}$. Le sous-ensemble des transactions utiles pour [1,3] est

[28] Les coefficients de cette matrice satisfont aux propriétés suivantes :
- composantes positives : $J_{p,q} \geq 0$
- éléments diagonaux : $J_{p,p} = 1$ pour p = 0, …, T
- les termes décroissent en ligne à partir des éléments diagonaux :

inclus dans l'ensemble des ventes répétées pertinentes pour [2,3] mais il n'en constitue pas la totalité. D'un point de vue informationnel, cela signifie que $I^{[1,3]} < I^2$. Le ratio $I^{[1,3]} / I^2$ mesure donc la part de l'information I^2 qui reste utile si l'on étend l'intervalle [2,3] par la gauche à l'intervalle [1,3]. D'une manière similaire, sur le coté droit $I^{[2,\ T]} / I^2$ représente la part de I^2 pertinente pour [0,T]. En conséquence, on appellera J la matrice de diffusion de l'information car elle mesure en fait la vitesse avec laquelle l'information se disperse.

A quoi correspondent maintenant les équations de la formule (18) ? Pour t = 2 on a :

$$(I^{[0,3]}/I^2)\, r_0 + (I^{[1,3]}/I^2)\, r_1 + 1\, r_2 + \ldots + (I^{[2,T]}/I^2)\, r_{T-1} = Y_2 = \tau^2\, \rho_2 \qquad (19)$$

A partir de la population des ventes répétées pertinentes pour [2,3], r_0 s'applique à la proportion de celles qui étaient déjà actives[29] pendant [0,1], r_1 à la proportion de celles qui étaient déjà actives pendant [1,2], … et r_T à la proportion de celles qui sont toujours en vie pour [T-1,T]. Tous ces rendements élémentaires produisent un rendement global qui est égal à $Y_2 = \tau^2\, \rho_2$. On aurait pu penser à priori que les pondérations auraient été réelles, mais la formule (19) indique que les cœfficients corrects sont des grandeurs informationnelles. Ceci semble indiquer que l'indice de ventes répétées $R = (r_0, r_1, \ldots, r_{T-1})$ est plutôt un concept d'ordre informationnel qu'un concept réel. La formule (18) s'interprète donc comme une décomposition informationnelle du rendement moyen, réalisé par les ventes répétées pertinentes pour un intervalle [t,t+1], en rendements élémentaires.

$1 \geq J_{p,\,p+1} \geq \ldots \geq J_{p,\,T}$ et $J_{p,\,0} \leq J_{p,\,1} \leq \ldots \leq 1$

[29] active = en vie

4. Un échantillon benchmark pour les ventes répétées

Avant d'exposer la reformulation générale[30] du RSI, nous développerons dans ce paragraphe une méthode de construction d'échantillons de référence pour les ventes répétées. Nous nous appuierons sur les concepts introduits précédemment et on continuera à se placer dans la situation simplifiée du paragraphe 3. Cet échantillon servira de base pour développer une méthodologie d'analyse de données dont on aura un premier aperçu dans le paragraphe 5 et qui sera appliquée en toute généralité dans le chapitre 2.

4.1. But et hypothèses

Une distribution des données de ventes répétées est fortement influencée par le contexte économique. Par exemple, si les prix immobiliers sont bas, les propriétaires qui souhaitent vendre leur bien peuvent décider de repousser leur transaction en espérant de meilleurs jours. Pour rendre les comparaisons entre différents contextes plus aisées, il pourrait être intéressant d'étudier un marché neutre dans lequel les événements économiques sont absents et d'en déduire quelle serait alors la distribution des ventes répétées ; en un mot de construire un échantillon benchmark.

Il faut bien sûr préciser ce que l'on entend par marché économiquement neutre. Ce concept sera associé aux hypothèses fictives suivantes :

- Le prix des biens ne varie jamais : $h_t = h_0$ pour tous les t
- La quantité globalement échangée sur le marché à toutes les dates est constante, on la notera K
- Les décisions d'achat et de revente sont indépendantes entre individus

[30] Cf. paragraphe 6.

- La longueur de la période de détention suit une loi exponentielle de paramètre $\lambda > 0$; λ ne dépend pas du propriétaire considéré

La dernière hypothèse est la plus importante. Elle implique que, conditionnellement à l'événement (achat à t = 0), la probabilité de ne pas avoir revendu le bien à l'instant t est égale à $e^{-\lambda t}$. Le choix d'une loi exponentielle est en fait assez irréaliste car il a pour conséquence la relation probabiliste suivante :

$$\text{Prob (revente} > t+1 \mid \text{revente} \geq t \text{) = Prob (revente} > s+1 \mid \text{revente} \geq s \text{)}$$

Cette formule indique que la probabilité de vendre le bien dans l'année à venir, sachant qu'il n'a pas été vendu jusqu'à présent, n'est pas influencée par l'ancienneté de l'achat[31]. D'une manière plus générale, si l'on introduit le taux de hasard[32] :

$$\lambda(t) = (1/\Delta t) * \text{Prob(revente} > t+\Delta t \mid \text{revente} \geq t)$$

qui mesure la probabilité de revente instantanée, on peut démontrer que choisir une distribution exponentielle revient en fait à choisir un taux de hasard constant. Concrètement cette modélisation signifie donc que l'intensité de l'activité de revente ne varie pas. Or, dans le monde réel il n'en est pas ainsi[33], même indépendamment du contexte économique. Pour un propriétaire courant on peut raisonnablement penser que le taux de hasard est d'abord faible (les reventes rapides sont rares), dans un second temps il s'accroît progressivement jusqu'à un certain palier (possiblement

[31] On parle de loi sans mémoire

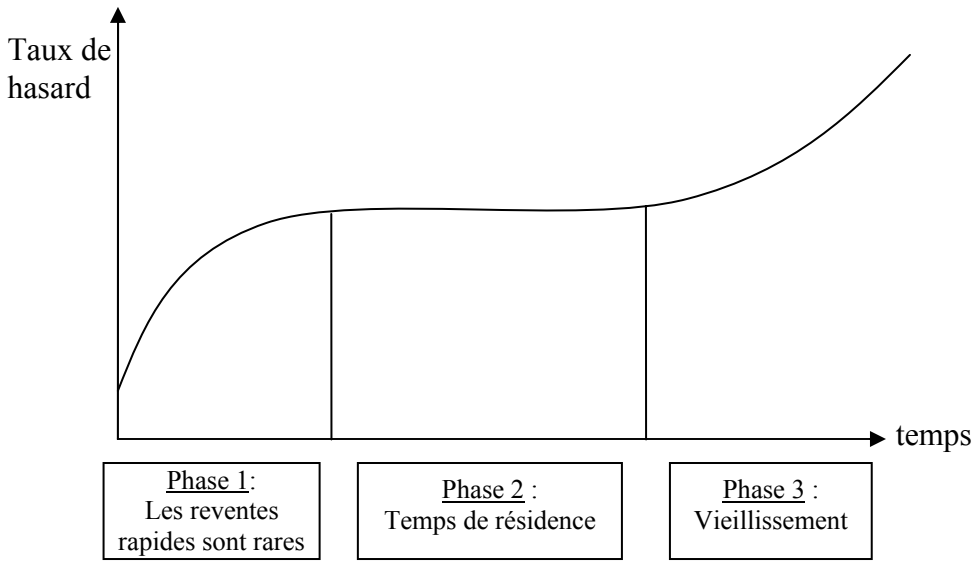
[32] Ce concept $\lambda(t)$ provient directement des modèles de survie, cf. Kalbfleisch, Prentice (2002) ou Cox, Oakes (1984). Il est par exemple utilisé dans les études économétriques portant sur la survenue des remboursements anticipés ou des défauts pour les mortgages (cf. Deng, Quigley, Van Order (2000)). Deville, Riva (2006) ont également employé cette approche pour étudier l'efficience du marché des options.

[33] Pour une modélisation moins naïve du phénomène de revente on pourra consulter l'article de Cheung, Yau, Hui (2004) qui, s'inspirant des études faites sur les remboursements anticipés des mortgages, régressent l'intensité du processus de revente sur les caractéristiques des biens.

modifié par le contexte économique), et enfin, dans une troisième phase (vieillissement), la possibilité d'un déménagement associé à la retraite ou encore la mort du propriétaire doivent probablement amener le taux à un niveau plus élevé. (cf. figure 4).

D'un point de vue comportemental, les réactions des propriétaires benchmark ne sont donc pas modifiées avec le temps. Ils garderont une humeur neutre et constante et seront complètement indifférents aux événements économiques ou personnels, d'autant plus que ces événements économiques ne se produiront pas d'après la première hypothèse de la modélisation ($h_t = h_0$). Le choix d'une distribution exponentielle est motivé par la volonté de construire un benchmark très simple où la décision de revente pourrait être comparée à la désintégration radioactive d'un atome. En effet, on sait d'après les résultats de la Physique nucléaire, que la probabilité d'un tel événement ne dépend pas de l'âge de la particule; en d'autres termes, le taux de désintégration $\lambda(t)$ est constant. Une des conséquences de cette situation est que l'on peut alors développer des modélisations radioactives très simples et très maniables. C'est cette maniabilité que l'on souhaite retrouver pour le benchmark des ventes répétées.

Figure 4: Taux de hasard hypothétique pour un propriétaire courant



4.2. La distribution des ventes répétées pour le benchmark

On déduit, à partir des hypothèses de modélisation du benchmark, la distribution des ventes répétées associée en fonction des deux paramètres du modèle : K et λ

4.2.1. Le taux de survie

A chaque date t, K transactions sont réalisées, les reventes correspondantes se produisent au cours de l'intervalle [t + 1 , $+\infty$ [. Comme les propriétaires agissent indépendamment et selon la même loi exponentielle, une application élémentaire de la loi des grands nombres, pour des valeurs de K suffisamment grandes, donne le pourcentage des K achats de la date t encore en vie aux dates t + 1, t + 2, … : $Ke^{-\lambda}$, $Ke^{-2\lambda}$, …

Tableau 4 : Evolution des K achats de la date t

	Transactions en vie au début de la période	Transactions en vie à la fin de la période	Nombre de reventes pendant la période
[t, t+1]	K	$K e^{-\lambda}$	$n_{t, t+1} = K (1 - e^{-\lambda})$
[t+1, t+2]	$K e^{-\lambda}$	$K e^{-2\lambda}$	$n_{t, t+2} = K e^{-\lambda} (1 - e^{-\lambda})$
[t+2, t+3]	$K e^{-2\lambda}$	$K e^{-3\lambda}$	$n_{t, t+3} = K e^{-2\lambda} (1 - e^{-\lambda})$
⋮			
[T-1, T]	$K e^{-(T-t-1)\lambda}$	$K e^{-(T-t)\lambda}$	$n_{t, T} = K e^{-(T-t-1)\lambda} (1 - e^{-\lambda})$

On note $\alpha = e^{-\lambda}$. On aura usuellement $\lambda \approx 0^{+}$ et donc $\alpha \approx 1^{-}$. Dans un ensemble de ventes répétées encore en vie[34] à une date donnée, α représente le pourcentage des survivants après un intervalle de temps élémentaire, il s'agit là du taux de survie

instantané. Après k unités de temps le taux de survie est α^k et le taux de disparition (de mortalité) est $d(k) = 1 - \alpha^k$. Pour le benchmark, d(k) ne dépend pas de la date d'achat. Les K ventes répétées initiées à la date t = 0 disparaîtront à la même vitesse que celles initiées à des dates plus récentes[35].

4.2.2. La distribution des $\{n_{i,j}\}$

Le nombre de ventes répétées avec un achat à t et une revente à t' est, d'après le tableau 4 :

$$n_{t,\,t'} = K\,(1/\alpha - 1)\,\alpha^{t'-t} = K'\,\alpha^{t'-t} \qquad (20)$$

avec $K' = K\,(1-\alpha)/\alpha$.

Le quotient $(1-\alpha)/\alpha$ réapparaîtra souvent dans les formules. Il peut s'interpréter facilement en utilisant l'approximation : $e^x \approx 1 + x$, pour x proche de 0.

$$(1-\alpha)/\alpha = (1-e^{-\lambda})/e^{-\lambda} = e^{\lambda} - 1 \approx \lambda \qquad (21)$$

Ce quotient correspond simplement au taux instantané de revente si λ est petit.[36] Si c'est le cas on a $K' \approx \lambda K$ et K' s'interprète alors comme le nombre de ventes répétées dans un ensemble de K qui disparaîtront au cours de la prochaine période de temps unitaire.

La distribution des $\{n_{i,j}\}$ qui découle de cette modélisation est présentée dans le tableau 5. Les valeurs de b_i (nombre de ventes répétées dans l'échantillon avec une

[34] La revente représente la mort.

[35] Dans une situation plus réaliste il en serait sans doute autrement. Il y aurait pour chaque date initiale une série de taux de disparition (d(t,k) for $t = t_i$, …, T-1) sans que les taux agissant aux mêmes dates soient égaux entre deux séries distinctes ; les différentes cohortes ayant des comportements différents.

[36] Si $\lambda = 0.1$, la valeur exacte pour le quotient $(1-\alpha)/\alpha$ est 0.105 ; l'approximation $(1-\alpha)/\alpha \approx \lambda$ produisant une erreur de 5%.

date d'achat à t_i) et s_j (nombre de ventes répétées dans l'échantillon avec une date de revente à t_j) peuvent être facilement calculées en remarquant que l'on a des progressions géométriques et en utilisant la relation $\alpha K' = K (1 - \alpha)$. De plus en sommant les b_i, ou les s_j , on obtient le nombre total de couples N dans l'échantillon en fonction de K, α et T.

On a : $$N = KT [1 - (\alpha / T (1 - \alpha)) * (1 - \alpha^T)] \qquad (22)$$

Tableau 5 : Distribution des ventes répétées pour l'échantillon benchmark

	0	1	2	…	t	t + 1	…	T	b_i
0		$K'\alpha$	$K'\alpha^2$		$K'\alpha^t$	$K'\alpha^{t+1}$		$K'\alpha^T$	$K (1 - \alpha^T)$
1			$K'\alpha$		$K'\alpha^{t-1}$	$K'\alpha^t$		$K'\alpha^{T-1}$	$K (1 - \alpha^{T-1})$
2					$K'\alpha^{t-2}$	$K'\alpha^{t-1}$		$K'\alpha^{T-2}$	$K (1 - \alpha^{T-2})$
¦									
t						$K'\alpha$		$K'\alpha^{T-t}$	$K (1 - \alpha^{T-t})$
t + 1								$K'\alpha^{T-t-1}$	$K (1 - \alpha^{T-t-1})$
¦									
T – 1								$K'\alpha$	$K (1 - \alpha)$
T									
s_j		$K(1-\alpha)$	$K(1-\alpha^2)$		$K(1-\alpha^t)$	$K(1-\alpha^{t+1})$		$K(1-\alpha^T)$	N

Les K biens négociés à une date "i" ne seront pas tous inclus dans l'échantillon des ventes répétées car certains d'entre eux ne seront revendus qu'après la date T. On aura donc toujours $b_i < K$ dans l'échantillon. Mais, évidemment plus T – i sera grand (c'est-à-dire plus il y aura de temps pour observer la revente), plus b_i sera proche de K.

Pendant [0, T], KT biens ont été échangés sur le marché. Pour la même raison que ci-dessus la taille de l'échantillon dans le tableau 5 sera nécessairement plus petite que ce nombre. La proportion manquante est une fonction de T et α qui se lit directement dans la formule (22), on la notera :

$$\pi = d(T) * (\alpha / T (1 - \alpha)) \approx d(T) / \lambda T \qquad (23)$$

(22) devient alors simplement : $N = K T (1 - \pi)$ (22')

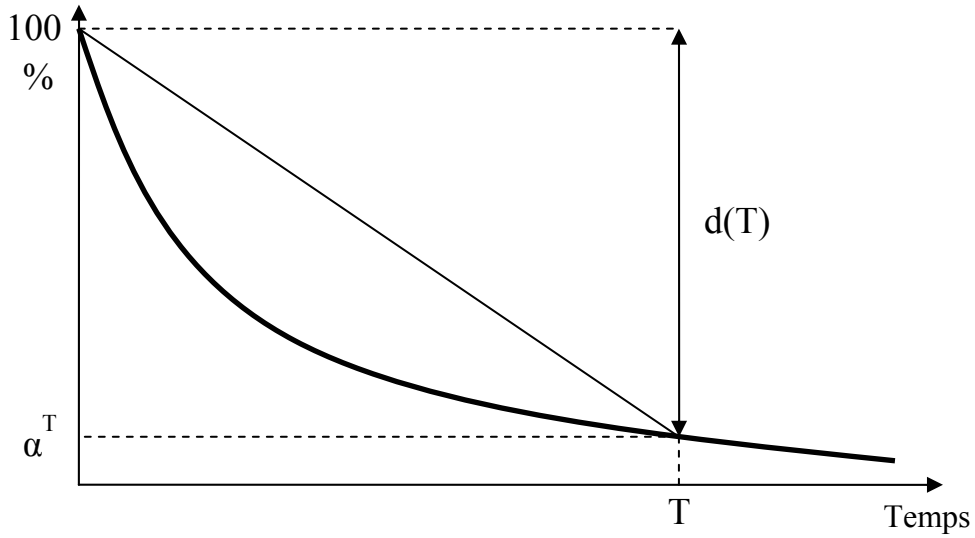
Il est facile de prouver que π décroît quand T augmente. L'interprétation de ce phénomène est évidente : la proportion manquante diminue quand le temps d'observation augmente. D'autre part si $\alpha = e^{-\lambda}$ augmente (ou en d'autres termes si les taux de survie d'une période à l'autre sont plus élevés) les reventes deviennent moins fréquentes et, très logiquement, la proportion manquante π devient plus importante.

4.2.3. Taux instantané et taux linéaire

Le quotient d(T) / T du pourcentage des biens ayant disparus après T unités de temps divisé par T s'analyse comme un taux linéaire de disparition. Dans la figure 5, on a représenté en traits épais la fonction de survie[37]. A t = 0, la proportion des biens encore en vie est de 100%, à T elle n'est plus que de α^T, les biens disparaissant à un taux multiplicatif constant. Si l'on veut obtenir le nombre moyen de disparitions[38] d'une période à l'autre et non plus un taux de disparition, on divisera le nombre de survivants d(T) par le nombre de périodes entre 0 et T.

[37] La fonction de survie donne à chaque date la proportion des transactions encore en vie, c'est-à-dire pour lesquelles la revente n'est pas encore intervenue.

Figure 5 : Taux instantané λ et taux linéaire $\lambda_{lin}(T)$



Le taux instantané λ de l'approche multiplicative est indépendant de T, tandis que le taux moyen linéaire λ_{lin} (T) = d(T)/ T est fortement dépendant de la durée de la période d'étude. Ces deux concepts de taux permettent de reformuler (23) en (23') ; la proportion manquante π correspond simplement au rapport entre le taux linéaire sur [0,T] et le taux instantané :

$$\pi \approx \lambda_{lin}(T) / \lambda \qquad (23')$$

4.3. La distribution de l'information pour le benchmark

A partir de la distribution des $\{n_{i,j}\}$ du tableau 5, on peut déduire la distribution informationnelle des $\{L_{i,j}\}$ en divisant par j – i (on se place toujours dans la situation simplifiée du paragraphe 3 où le bruit est mesuré par (j – i)). Les sommes en lignes et en colonnes des $L_{i,j}$ ne sont plus aussi simples que dans le tableau 5 car la progression n'est plus géométrique. On introduit alors la quantité u_n définie par :

[38] En pourcentage de l'effectif de départ

$$u_n = \alpha + \alpha^2 / 2 + \alpha^3 / 3 + \ldots + \alpha^n / n \qquad (24)$$

Les diverses grandeurs informationnelles s'exprimeront toutes à partir de u_n, comme on peut déjà le constater pour les B_i et les S_j (tableau 6). Ce nombre u_n correspond en fait à la quantité d'information que fournit un échantillon de taille $K = 1 / \lambda$ pendant n unités temporelles[39]. La limite de $(u_n)_{n \in N}$ est $\ell = -\ln(1 - \alpha)$. Malheureusement comme $\alpha \approx 1^-$ la vitesse de convergence est très faible, et on ne peut pas raisonnablement utiliser la limite comme une approximation de u_n[40].

Tableau 6 : Distribution des $\{L_{i,j}\}$ pour le benchmark exponentiel

	0	1	2	…	t	t + 1	…	T	B_i
0		K'α	$K'\alpha^2/2$		$K'\alpha^t / t$	$K'\alpha^{t+1}/t+1$		$K'\alpha^T / T$	$K' u_T$
1			K'α		$K'\alpha^{t-1}/t-1$	$K'\alpha^t/t$		$K'\alpha^{T-1}/T-1$	$K' u_{T-1}$
2					$K'\alpha^{t-2}/t-2$	$K'\alpha^{t-1}/t-1$		$K'\alpha^{T-2}/T-2$	$K' u_{T-2}$
¦									
t						K'α		$K'\alpha^{T-t}/T-t$	$K' u_{T-t}$
t + 1								$K'\alpha^{T-t-1}/T-t-1$	$K' u_{T-t-1}$
¦									
T – 1								K'α	$K' u_1$
T									
S_j		$K' u_1$	$K' u_2$		$K' u_t$	$K' u_{t+1}$		$K' u_T$	I

Sur le plan des notations on veillera à ne pas confondre la quantité d'information B_i, fournie par les ventes répétées ayant une date d'achat à t_i, et B^i la quantité d'information pertinente pour l'intervalle [i, i+1], vue du côté achat. De même on

[39] Pour $K = 1/\lambda$ on a $K' = \lambda K = \lambda / \lambda = 1$. En reprenant la première ligne du tableau 6 la quantité d'information délivrée, pendant n unités de temps, par ces $(1/\lambda)$ transactions initiées à t = 0 est alors bien égale à u_n

distinguera la quantité d'information S_j, fournie par les ventes répétées ayant une date de revente à t_j, et S^j la quantité d'information pertinente pour l'intervalle [j, j+1], vue du côté revente. On aura ainsi $B^t = S^t = I^t$, mais en général la relation $B_t = S_t$ sera fausse.

La quantité d'information globale I est :

$$\begin{aligned} I &= \sum_{i = 0,\ldots,T-1} B_i = K' \sum_{i = 1,\ldots,T} u_i = K' \sum_{i = 1,\ldots,T} \sum_{j = 1,\ldots,i} \alpha^j / j \\ &= K' \sum_{k = 1,\ldots,T} [(T - k + 1) / k] \alpha^k \qquad \text{(somme sur les diagonales)} \\ &= K' (T + 1) \sum_{k = 1,\ldots,T} \alpha^k / k - K' \sum_{k = 1,\ldots,T} \alpha^k \\ &= K' (T + 1) u_T - K' (\alpha /(1 - \alpha)) * (1 - \alpha^T) \\ &= K' (T + 1) u_T - K' T [(\alpha / T(1 - \alpha)) * (1 - \alpha^T)] \end{aligned}$$

On reconnaît la proportion manquante π de la relation (23), la formule donnant la quantité d'information totale s'écrira donc :

$$I = K' T [(1 + 1 / T) u_T - \pi] \qquad (25)$$

Pour le benchmark, l'information totale est donc une fonction de la proportion manquante π et de u_T. Comme ces deux quantités dépendent de α et T, il en est de même pour I. Cette expression permet de prouver facilement que I croît quand l'horizon d'observation T augmente et quand le taux de survie α décroît.

Pour poursuivre l'analyse de (25) il serait intéressant de connaître la quantité d'information totale produite par les KT biens échangés sur le marché[41] entre 0 et T et de déterminer la proportion révélée dans l'échantillon de ventes répétées.

---

[40] Par exemple, si $\alpha = 0.95$ on a $\ell \approx 2.9957$, $u_{10} \approx 2.4805$, $u_{20} \approx 2.7944$. Si l'unité de temps est l'année après 20 ans l'erreur serait encore de 6.7%.

[41] Indépendamment de la date de revente.

A t = 0 K propriétés arrivent sur le marché et, d'après le tableau 5, $K(1 - \alpha^T)$ d'entre elles ont été revendues avant T, fournissant ainsi une quantité d'information de K' u_T (tableau 6). Si l'on n'avait pas mis de barrière à la date T, tous les biens auraient été revendues et l'information produite aurait été de K'ℓ. Ce raisonnement s'appliquant à toutes les propriétés pour lesquelles l'achat a été réalisé entre 0 et T, la quantité asymptotique d'information est donc de K'ℓT.

Puisque $\ell = -\ln(1-\alpha)$, la proportion d'information révélée ψ est alors :

$$\psi = I / (K'\ell\, T) = [(1 + 1 / T) u_T - \pi] / (-\ln(1-\alpha)) \qquad (26)$$

L'expression entre crochets tend vers $-\ln(1-\alpha)$ avec T et la limite de ψ, quand $T \rightarrow +\infty$, est évidemment égale à 1. Quand l'horizon d'observation est infini on observe toutes les reventes et l'information est complètement révélée.

4.4. Effectifs et quantités d'information pertinents pour un intervalle [t',t+1]

4.4.1. Les effectifs pertinents

Grâce aux tableaux 2 et 5, on peut établir les formules des effectifs pertinents pour t = 0, 1, …, T-1 et $t' \leq t$:

$$n^t = (K / (1 - \alpha)) * d (T - t) d(t + 1) \qquad (27)$$

$$n^{[t',t+1]} = (K /(1 - \alpha)) * (1 - d (t - t')) d(T - t) d(t' + 1) \qquad (27')$$

$n^t = \sum_{i \leq t < j} n_{i,j}$ dénombre les effectifs des ventes répétées pertinentes pour [t, t+1], tandis que $n^{[t',t+1]} = \sum_{i \leq t' \leq t < j} n_{i,j}$ compte les couples pertinents pour [t', t+1]. $n^{[t',t+1]}$ est ainsi à n^t ce que $I^{[t',t+1]}$ est à I^t.

4.4.2. La quantité d'information $I^{[t',t+1]}$

Le calcul de l'information pertinente, $I^{[t',t+1]}$ pour t' < t, est un peu plus compliqué mais la difficulté peut être surmontée en raisonnant graphiquement. On se reportera à l'annexe 3 pour les détails du raisonnement.

Le résultat s'écrit :

$$I^{[t',t+1]} = K' \mathcal{U}(t', t, T) + K\, \alpha^{t-t'-1} (1 - \alpha^{t'+1}) (1 - \alpha^{T-t}) \qquad (28)$$

Où $\mathcal{U}$ est une fonction définie pour $0 \leq t' < t < T$ par :

$$\mathcal{U}(t',t,T) = U(t,T) - (t\, U(t-t'-1,t) - t'\, U(t-t'-1,T-t'-1)) + T\, U(T-t'-1,T)$$

Avec : $U(m,n) = u_n - u_m = \alpha^{m+1} / (m+1) + \ldots + \alpha^n / n$ pour $m \leq n$

Et : $u_0 = 0$ par convention

Il faut remarquer ici que les expressions de U(m,n) et $\mathcal{U}(t',t, T)$ sont toutes bâties à partir de la quantité de base u_n.

Dans la deuxième partie de $I^{[t',t+1]}$ on reconnaît une expression très voisine de $n^{[t',t+1]}$. En utilisant (27') on obtient :

$$I^{[t',t+1]} = ((1 - \alpha) / \alpha) * (n^{[t', t+1]} + K\, \mathcal{U}(t', t, T)) \qquad (28')$$

$$I^{[t',t+1]} \approx \lambda (n^{[t', t+1]} + K \mathcal{U}(t', t, T)) \qquad (28'')$$

4.4.3. La quantité d'information I^t

Pour $I^{[t, t+1]} = I^t$, les formules ci-dessus ne s'appliquent plus car la démonstration doit être légèrement adaptée (annexe 4). On obtient, en introduisant la convention de calcul $u_{-1} = 0$ et en conservant les mêmes notations :

$$I^t = K' \mathcal{U}(t, t, T) + (K / \alpha) [(1 - \alpha^{t+1}) (1 - \alpha^{T-t}) - (1 - \alpha)] \qquad (29)$$

Dans la deuxième partie la quantité n^t apparaît via la formule (27), on a alors :

$$I^t = ((1 - \alpha) / \alpha) * (n^t + K(\mathcal{U}(t, t, T) - 1)) \qquad (29')$$

Ou encore : $I^t \approx \lambda (n^t + K (\mathcal{U}(t, t, T) - 1))$ (29'')

On peut vérifier de visu que la formule pour t' = t n'est effectivement pas une simple généralisation de (28') ou (28'').

4.5. Taux, fréquences et périodes, matrices d'information

Les grandeurs R et P ne sont pas utiles pour les échantillons de référence. H restant constant à toutes les dates, les moyennes $H_f(t)$ et $H_p(t)$ sont également constantes et elles valent toutes h_0. La formule (14') implique alors $\rho_t = 0$ pour tous les temps et les formules (17) et (18) donnent, de même, $r_t = 0$. Ce résultat est assez évident intuitivement car un taux, quelle que soit sa définition, ne peut pas différer de zéro si les prix restent constants.

Un exemple de matrice informationnelle pour le benchmark sera fournit dans le paragraphe suivant pour T = 5. Il ne reste plus maintenant qu'à présenter les formules des fréquences et des périodes pour achever l'étude de l'échantillon benchmark.

La fréquence moyenne de détention F = I / N pour l'échantillon benchmark pris dans son ensemble (cf. annexe 1) peut être calculée avec les formules (22') et (25).

On a : $F = ((1 - \alpha) / \alpha) * [((1 + 1 / T) u_T - \pi) / (1 - \pi)]$ (30)

D'où : $F \approx [(1 + 1/ T) \lambda u_T - \lambda_{lin}(T)] / [1 - \lambda_{lin}(T) / \lambda)]$ (30')

Pour la population des ventes répétées pertinentes pour l'intervalle [t, t +1], on peut utiliser les formules (29') ou (29'') :

Comme $F^t = I^t / n^t$

On a : $F^t = ((1 - \alpha) / \alpha) * (1 + (K/ n^t)*(\mathcal{U}(t, t, T) - 1))$ (31)

D'où : $F^t \approx \lambda * (1 + (K/ n^t)*(\mathcal{U}(t, t, T) - 1))$ (31')

De (30) ou (30') on peut déduire la limite de F quand $T \rightarrow +\infty$:

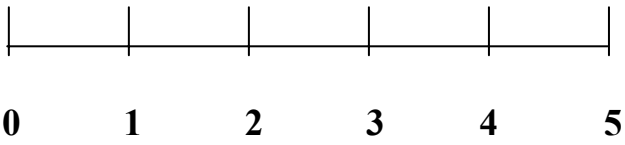
$$\lim_{T\rightarrow+\infty} F = ((1 - \alpha) / \alpha) \ell \approx \lambda \ell \quad \text{d'où} \quad \lim_{T\rightarrow+\infty} \tau = 1 / \lambda\ell \qquad (32)$$

Quand $T \rightarrow +\infty$ la proportion manquante se réduit. Or, comme le temps entre l'achat et la revente suit une loi exponentielle, on aurait pu penser que la limite de la période de détention moyenne τ aurait été $1/ \lambda$ et pas $1 / \lambda \ell$ (en effet, si X suit une distribution exponentielle $\mathcal{E}(\lambda)$, on a $E(X) = 1/ \lambda$). Ce résultat s'explique en remarquant que, dans le modèle des ventes répétées, la moyenne se calcule d'abord avec les fréquences et ce n'est qu'ensuite que l'on inverse la valeur obtenue ; τ est donc une moyenne harmonique. Or, une espérance est toujours une moyenne arithmétique. Le vrai calcul[42] à effectuer pour retrouver τ n'est pas E(X) mais $[E(1/X)]^{-1}$ et il n'y a pas de raison particulière qui pourrait amener à penser que ces deux grandeurs sont égales. Le nombre ℓ qui apparaît ici peut s'interpréter comme un facteur de distorsion entre la structure arithmétique et la structure harmonique.

5. Un exemple élémentaire d'application du benchmark

Cette partie applique les concepts développés dans le paragraphe 4 à une situation simple. Comme on le verra dans le chapitre 2, l'échantillon benchmark servira de support à une méthodologie d'analyse des données réelles ; le présent paragraphe donne un premier aperçu de cette méthode.

On supposera ici qu'il n'y a que cinq intervalles de temps.



On commencera par se donner un scénario économique et un échantillon de ventes répétées associé. Comme seules ces données seront observables en pratique, le but de la méthodologie sera alors de retrouver les phénomènes économiques les ayant générées. Pour atteindre cet objectif il faudra étalonner le benchmark, calculer ensuite les grandeurs associées aux différentes distributions puis mettre en œuvre la méthode d'analyse de données, dans le dernier paragraphe.

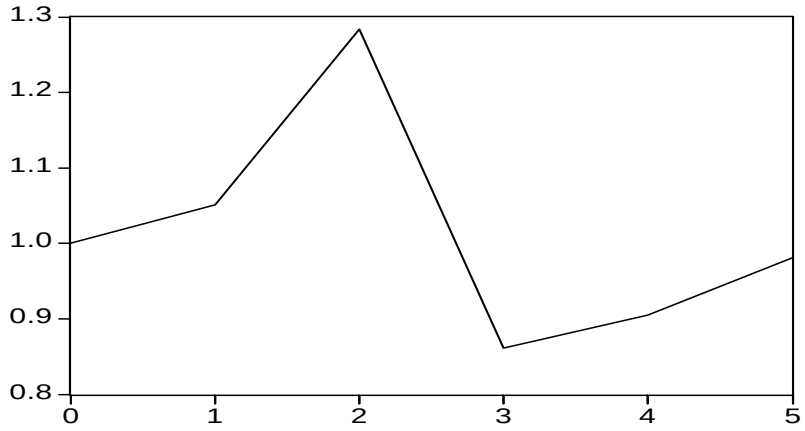
5.1. Données et scénario

Les valeurs de l'indice de prix H sont présentées dans la figure 6 et illustrées graphiquement. Elles correspondent à des taux de croissance continus de 5% pour la première période, 20% pour la seconde, - 40% pour la troisième, 5% et 8% pour la quatrième et la cinquième.

[42] 1/X est la fréquence associée à chaque couple, E(1/X) la moyenne des fréquence et son inverse est

Figure 6 : Valeurs de l'indice de prix et représentation graphique

	0	1	2	3	4	5
H	1	1.051	1.284	0.861	0.905	0.980



Le choix de ces valeurs pour H est motivé par le scénario suivant :
La première année la croissance des prix immobiliers est assez classique (5%). La deuxième année elle atteint 20%, on pourra interpréter ce phénomène comme l'apparition d'une bulle. Immédiatement après les prix s'effondrent brutalement (- 40%). Pendant les deux dernières périodes on retrouve un niveau standard de croissance (5% et 8%).

Ce scénario est bien sur une caricature. Le but que l'on poursuit ici consiste simplement à illustrer et à tester le modèle sur une situation basique, où les mouvements de prix sont nets et marqués.

La distribution des $\{n_{i,j}\}$ de l'échantillon d'estimation est présentée dans le tableau 7. Quatre phénomènes fondamentaux ont motivé les choix de ces valeurs :

---

donc la moyenne harmonique des périodes de détention.

- $s_2 = 21$: conséquence du fort taux de croissance de la deuxième période ; les propriétaires essayent de profiter de ces prix élevés pour réaliser leur investissement.
- $s_3 = 9$: ce chiffre témoigne au contraire d'un niveau de revente très bas associé à la chute brutale des valeurs de l'indice ; les vendeurs potentiels préfèrent attendre de meilleurs jours pour effectuer leur transaction.
- $s_4 = 30$ et $s_5 = 32$: le marché se reprend (en volume) ; les propriétaires effrayés par le krach en profitent pour vendre massivement.
- $b_3 = 16$: les prix étant assez bas à t = 3, certains investisseurs pensent qu'il s'agit du bon moment pour acheter.

Tableau 7 : Distribution des $\{n_{i,j}\}$

t_i \ t_j	0	1	2	3	4	5	b_i : achats cumulés à t_i
0		8	11	2	7	6	34
1			10	3	8	4	25
2				4	8	4	16
3					7	9	16
4						9	9
5							
s_j : ventes cumulées à		8	21	9	30	32	100

Il faut bien comprendre que pour le scénario et les phénomènes évoqués ci-dessus on se place dans une situation omnisciente où l'on connaît tous les rouages de l'économie. Pour l'économètre, qui ne voit que les conséquences de la situation fondamentale par l'intermédiaire des données, la position est différente.

Une analyse directe des $\{n_{i,j}\}$ est ainsi relativement difficile pour lui, car il ne peut pas réellement comparer s_1 et s_5 (par exemple). s_5 correspond en effet aux couples

pour lesquels l'achat a été effectué à t = 0, tandis que pour s_1 l'acte d'achat a pu être réalisé entre t = 0 et t = 4. Il est donc normal d'avoir $s_1 < s_5$. D'une manière plus générale, les $\{s_j\}$ sont en fait globalement croissants et les $\{b_i\}$ globalement décroissants. Par lecture directe, on pourrait éventuellement remarquer que $b_2 = 16$ et $b_3 = 16$ doivent être symptomatiques d'une situation anormale, mais comme on ne connaît pas le niveau de référence de ces grandeurs il est difficile de trancher entre un b_2 trop faible et un b_3 trop élevé.

Une analyse de la distribution des $\{n_{i,j}\}$ ne peut donc pas procéder d'une comparaison entre les niveaux des $\{b_i\}$ et des $\{s_j\}$ à deux dates différentes. Pour étudier, de façon pertinente, les fluctuations de ces grandeurs il faut pouvoir déterminer un niveau « normal » et y rapporter les données de l'échantillon réel. Cette fonction sera remplie par l'échantillon benchmark qui, étant neutre par construction, permettra de mettre en évidence les phénomènes non standards (par exemple une augmentation soudaine du niveau des reventes). Notre méthodologie d'analyse de données s'appuiera sur des indicateurs interprétables directement, qui ne présenteront pas les difficultés et les ambiguïtés d'une analyse directe des $\{n_{i,j}\}$ et qui permettront de retrouver les quatre phénomènes décrits explicitement ci-dessus[43].

A partir de la distribution des $\{n_{i,j}\}$, on déduit la distribution informationnelle des $\{L_{i,j}\}$ présentée dans le tableau 8. La différence la plus notable entre ces deux distributions porte sur l'évolution des valeurs au fur et à mesure que l'on s'éloigne de la diagonale. Dans le tableau 7, le coin supérieur droit comporte quelques valeurs élevées ($n_{0,4} = 7$; $n_{0,5} = 6$) mais ce n'est plus le cas lorsque l'on passe au tableau 8 ($L_{0,4} = 1.75$; $L_{0,5} = 1.2$). Une cellule éloignée de la diagonale est moins informative qu'une cellule proche de cette ligne ; sa contribution au calcul d'un indice de ventes répétées sera donc atténuée.

Les N = 100 ventes répétées fournissent une niveau d'information de I = 62.12. La fréquence moyenne est F = I / N = 0.6212 et la moyenne harmonique des périodes de détention est $\tau = N / I \approx 1.610$.

Tableau 8 : Distribution des $L_{i,j} = n_{i,j} / (j - i)$

t_j \ t_i	0	1	2	3	4	5	B_i : information fournie par les achats à t_i
0		8	5,5	0,67	1,75	1,2	17,12
1			10	1,5	2,67	1	15,17
2				4	4	1,33	9,33
3					7	4,5	11,5
4						9	9
5							
S_j : information fournie par les reventes à t_j		8	15,5	6,17	15,42	17	62,12

5.2. L'étalonnage du benchmark

Pour rapporter correctement un échantillon réel à un échantillon benchmark il faut pouvoir assurer une certaine homogénéité (dans la taille par exemple) entre les deux lots de données ; c'est la fonction du procédé d'étalonnage présenté dans ce paragraphe.

Pour une maturité T fixée, le benchmark est complètement déterminé par la donnée du couple (K, λ) ; plusieurs choix sont toutefois possibles. Celui que l'on

[43] Pour T = 5 l'analyse directe des $\{n_{i,j}\}$ est, à la rigueur, encore envisageable mais pour des échantillons réels (T = 40 par exemple) elle deviendrait inextricable. Il faut donc construire une procédure d'analyse efficace, susceptible de gérer la complexité des grands échantillons.

privilégiera ici est motivé par la volonté de conserver les agrégats (I, N) aux mêmes niveaux ou, de manière équivalente, les agrégats (N, F).

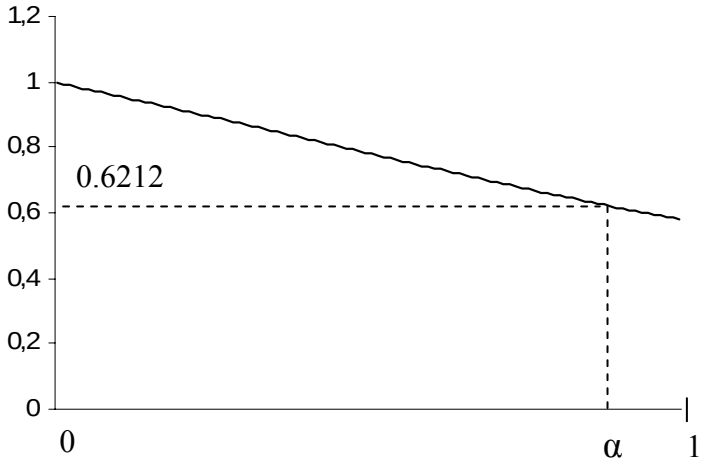
Les formules (30) et (23) donnent la fréquence pour l'échantillon benchmark. Ces expressions ne dépendent pas de K, on peut donc en déduire la valeur de α en résolvant l'équation F = 0.6112. Ce problème étant relativement complexe, une formule fermée n'est pas évidente à trouver, on se contentera d'une résolution par des méthodes numériques.

λ variant dans $]0, +\infty[$, α variera dans $]0,1[$ puisque $\alpha = e^{-\lambda}$. Pour T = 5, la fréquence est une fonction de α dont la représentation graphique est donnée dans la figure 7. Cette fonction est décroissante et en utilisant la méthode dichotomique on obtient $\alpha \approx 0.88824$; ce qui correspond à $\lambda \approx 0.11851$. Une fois la valeur de α connue, la formule (22') permet de déterminer la valeur de K en résolvant l'équation N = 100. Or, comme π ne dépend pas de K, la solution est immédiate :

$$K = N / (T*(1 - \pi))$$

Dans le cas qui nous occupe, on trouve ainsi $K \approx 69.127$. Les trois paramètres qui déterminent entièrement le benchmark sont donc T = 5, K = 69.127 et $\alpha = 0.88824$.

Figure 7 : F fonction de α pour T = 5



Cette technique d'étalonnage amène deux remarques. Premièrement le choix des équations de calibration est libre. Celui effectué ici se fondait sur les formules (30) pour F et (22') pour N car on a souhaité conserver les mêmes valeurs pour ces agrégats. Ce choix n'est pas impératif et d'autres stratégies auraient pu être utilisées. Par exemple si l'on décide que le comportement de référence est celui des vendeurs au début de l'intervalle [0,T], on peut utiliser les formules portant sur s_1 et s_2 pour en inférer les valeurs de K et α (tableau 5). Ou, d'une manière plus générale, on peut utiliser toutes les valeurs de s_1 à s_T et étalonner K et α de sorte que les valeurs théoriques du benchmark $s_1(K, \alpha)$, …, $s_T(K, \alpha)$ soient le plus proche possible (en un sens à définir) de la série des valeurs réelles. On prend alors comme référence le comportement des vendeurs sur l'intervalle [0,T]. Un étalonnage côté achat est aussi possible en partant de la série des $\{b_i\}$ et l'on peut également réaliser des calibrations à l'aide des séries informationnelles $\{B_i\}$ et $\{S_j\}$. Comme on le voit, le processus d'étalonnage présente des degrés de liberté intéressants permettant d'adapter le modèle aux problématiques économiques.

En second lieu, il faut remarquer que la calibration n'est pas toujours réalisable. Dans l'exemple ci-dessus on a pu déduire de l'équation F = 0.6212 la valeur de α par une méthode graphique illustrée dans la figure 7. Il semble cependant que certaines valeurs de F ne soient pas atteignables par le benchmark exponentiel. Seules les fréquences entre 0.58 (approximation graphique) et 1 peuvent ainsi être reproduites en choisissant un α idoine dans l'intervalle]0,1[(pour T = 5). Or, malheureusement, on peut rencontrer des distributions réelles présentant des fréquences plus basses.

Ainsi, dans la situation du tableau 9, N vaut 100 et I vaut 55.45, on a donc F = I / N = 0.5545. Ce niveau de fréquence ne peut pas être atteint si l'on modélise le benchmark avec une distribution exponentielle pour la loi de revente. Ce phénomène est probablement la conséquence des valeurs très élevées que l'on trouve dans les deuxièmes colonnes. Cet échantillon, en s'éloignant trop du comportement lisse de la fonction exponentielle, rend l'étalonnage irréalisable. La solution consisterait alors à

choisir une autre loi pour la durée de détention et à reprendre les calculs des différentes grandeurs du benchmark.

Tableau 9 : Un exemple de distribution inatteignable par un benchmark exponentiel (T = 5)

$n_{i,j}$	0	1	2	3	4	5	b_i		$L_{i,j}$	0	1	2	3	4	5	B_i
0		5	20	1	10	6	42		0		5	10	0,33	2,5	1,2	19,03
1			10	2	10	5	27		1			10	1	3,33	1,25	15,58
2				1	10	4	15		2				1	5	1,33	7,33
3					8	5	13		3					8	2,5	10,5
4						3	3		4						3	3
5									5							
s_j		5	30	4	38	23	100		S_j		5	20	2,33	18,83	9,28	55,45

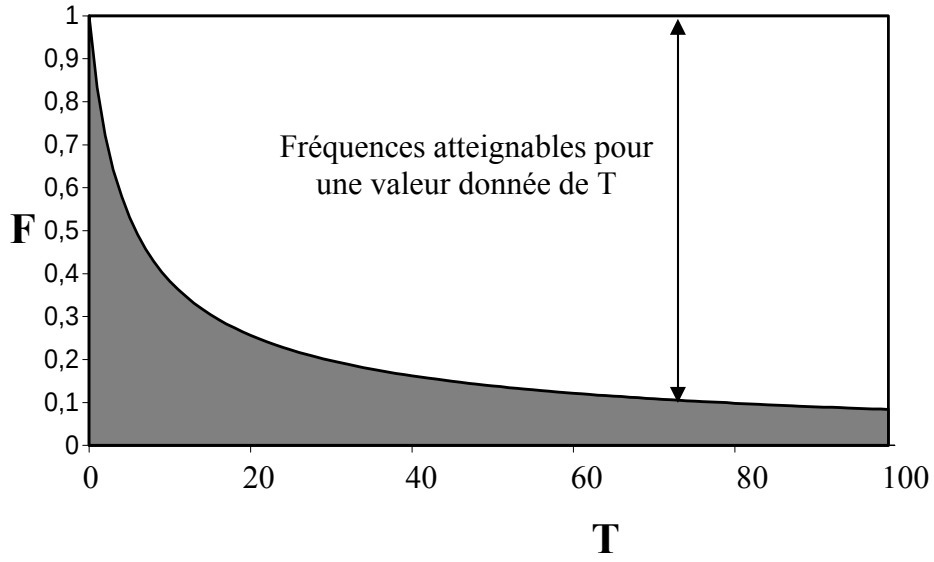
Deux questions naturelles se posent alors :

- Pour un T fixé, quelles sont les fréquences reproductibles par un benchmark exponentiel ?
- La fonction F(α) est-elle toujours décroissante sur]0,1[?

Cette seconde question est également d'importance car si la réponse est négative, il existerait dans certains cas plusieurs possibilités pour le paramètre α. Une étude basique de la fonction F (cf. annexe 5) permet d'y répondre. On y établit que la fonction F(α), pour une valeur fixée de T, décroît sur l'intervalle]0,1[de 1 à $F(1^-)$. Toutes les valeurs situées entre ces deux bornes sont donc atteignables par un benchmark exponentiel. Cette aire des possibles est représentée dans la figure 8. La limite inférieure vaut :

$$F(1^-) = 2\,[(1 + 1 / T)\, u_T(1) - 1\,] / (\,1 + T) \quad \text{avec} \quad u_T(1) = 1 + 1/2 + 1/3 + \ldots + 1/T$$

Figure 8 : Fréquences atteignables par un benchmark exponentiel en fonction de T



5.3. Les variables du benchmark

L'étape d'étalonnage a donné, α = 0,88824 et K = 69,127. Les distributions correspondantes des $\{n_{i,j}\}$ et des $\{L_{i,j}\}$ du benchmark sont représentées ci-dessous.

Tableau 10 : Distributions pour le benchmark étalonné

$n_{i,j}$	0	1	2	3	4	5	b_i		$L_{i,j}$	0	1	2	3	4	5	B_i
0		7,73	6,86	6,10	5,41	4,81	30,91		0		7,73	3,43	2,03	1,35	0.96	15,50
1			7,73	6,86	6,10	5,41	26,10		1			7,73	3,43	2,03	1,35	14,54
2				7,73	6,86	6,10	20,68		2				7,73	3,43	2,03	13.19
3					7,73	6,86	14,59		3					7,73	3,43	11,16
4						7,73	7,73		4						7,73	7,73
5							100		5							
s_j		7,73	14,59	20,68	26,10	30,91	100,01		S_j		7,73	11,16	13,19	14,54	15,50	62,12

Pour l'échantillon réel on a N = 100 et I = 62,116667 (tableaux 7 et 8). Les valeurs du tableau 10 ont été arrondies, mais si l'on augmente le degré de précision des calculs on atteint N = 100,0001 et I = 62,11674 sans difficultés. Les deux agrégats N et I sont donc bien à leurs niveaux cibles ; les valeurs correspondantes pour F et τ sont alors respectivement de 0.6212 et 1.610.

Les équivalents temporellement variables de N, I, F et τ sont fournis par le tableau suivant en appliquant les formules (27), (29') et (31).

Tableau 11 : n^t, I^t, F^t, τ^t pour le benchmark

t	0	1	2	3	4
n^t	30,91	49,28	55,37	49,28	30,91
I^t	15,50	22,32	24,35	22,32	15,50
F^t	0,50	0,45	0,44	0,45	0,50
τ^t	1,99	2,21	2,27	2,21	1,99

Les matrices informationnelles sont :

$$\hat{I} = \begin{pmatrix} 15,50 & 7,78 & 4,35 & 2,32 & 0,96 \\ 7,78 & 22,3 & 11,16 & 5,70 & 2,32 \\ 4,35 & 11,63 & 24,35 & 11,16 & 4,35 \\ 2,32 & 5,70 & 11,16 & 22,32 & 7,78 \\ 0,96 & 2,32 & 4,35 & 7,78 & 15,50 \end{pmatrix}$$

$$J = \begin{pmatrix} 1 & 0.50 & 0.28 & 0.15 & 0.06 \\ 0.35 & 1 & 0.50 & 0.26 & 0.10 \\ 0.18 & 0.46 & 1 & 0.46 & 0.18 \\ 0.10 & 0.26 & 0.50 & 1 & 0.35 \\ 0.06 & 0.15 & 0.28 & 0.50 & 1 \end{pmatrix}$$

On peut remarquer que la structure régulière du benchmark produit des résultats symétriques[44] (par exemple $F^0 = F^4$). On sait qu'il existe pour la matrice Î une propriété de symétrie forte, celle-ci découlant de la définition même de la matrice, mais que la matrice J ne présente pas, dans le cas général, une telle caractéristique. Pour le benchmark exponentiel il en est toujours ainsi, mais J possède toutefois une propriété de symétrie faible : si l'on inverse la $i^{ème}$ ligne on retrouve la $(T - i)^{ème}$ ligne. Cette caractéristique est une conséquence directe de la définition du benchmark et n'est pas vérifiée pour les échantillons réels.

Enfin, le taux linéaire équivalent à λ est $\lambda_{lin}(5) = 0.08942$, et (26) fournit la proportion d'information révélée au cours de l'intervalle [0,T] : $\psi = 65.18\%$.

5.4. Les variables de l'échantillon réel

On sait déjà que : $N = 100$ $I = 62.12$ $F = 0.6212$ $\tau \approx 1.610$.

Pour la distribution réelle n^t, I^t, F^t et τ^t ne peuvent pas se calculer par une formule simple comme cela a été possible pour le benchmark ((27), (29') ou (31)) car il

n'existe plus de régularité explicite. Le seul moyen consiste à utiliser les définitions basiques de ces concepts. On obtient alors les résultats suivants :

Tableau 12 : n^t, I^t, F^t, τ^t pour la distribution réelle

t	0	1	2	3	4
n^t	34	51	46	53	32
I^t	17,12	24,28	18,12	23,45	17,03
F^t	0,50	0,48	0,39	0,44	0,53
τ^t	1,99	2,10	2,54	2,26	1,88

Les matrices informationnelles sont :

$$\hat{I} = \begin{pmatrix} 17,12 & 9,12 & 3,62 & 2,95 & 1,20 \\ 9,12 & 24,28 & 8,78 & 6,62 & 2,20 \\ 3,62 & 8,78 & 18,12 & 11,95 & 3,53 \\ 2,95 & 6,62 & 11,95 & 23,45 & 8,03 \\ 1,20 & 2,20 & 3,53 & 8,03 & 17,03 \end{pmatrix}$$

$$J = \begin{pmatrix} 1 & 0,53 & 0,21 & 0,17 & 0,07 \\ 0,38 & 1 & 0,36 & 0,27 & 0,09 \\ 0,20 & 0,48 & 1 & 0,66 & 0,20 \\ 0,13 & 0,28 & 0,51 & 1 & 0,34 \\ 0,07 & 0,13 & 0,21 & 0,47 & 1 \end{pmatrix}$$

[44] Cette symétrie provient directement des valeurs du tableau 10 qui sont symétriques par rapport à la deuxième diagonale (/)

Les valeurs nécessaires à l'application de la formule (14') pour P sont fournies par le tableau 13.

Tableau 13 : $H_p(t)$, $H_f(t)$, ρ_t

t	0	1	2	3	4
$H_p(t)$	1	1,03	1,15	0,99	0,93
$H_f(t)$	1,09	1,13	0,90	0,93	0,98
ρ_t	0,043	0,045	-0,096	-0,028	0,026

Les produits matriciels (17) ou (18) donnent le vecteur R, dont on déduit les valeurs cumulées Ind (cf. tableau 14). Et, comme mentionné précédemment, on a bien dans ce cas particulier Ind = H. Les deux indices sont égaux dans ce contexte simplifié.

Tableau 14 : Valeurs du RSI

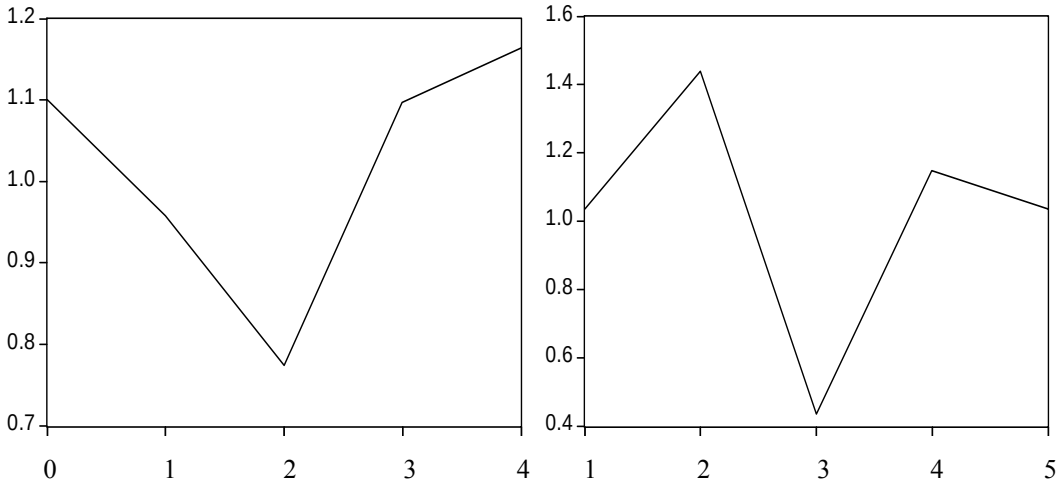
	0	1	2	3	4	5
r_t	0.0497	0.2002	-0.3996	0.0498	0.0796	
Ind_t	1	1.051	1.284	0.861	0.905	0.980

5.5. L'analyse de données

Les calculs précédents permettent maintenant de mettre en œuvre l'analyse de données. L'activité du marché peut être étudiée avec la figure 9 en divisant les s_j réels du tableau 7 par les s_j correspondants pour le benchmark (tableau 10) ; de même pour les b_i, on divise les quantités réelles par leurs équivalents benchmark. On obtient alors des indicateurs immédiatement exploitables et ne présentant pas les ambiguïtés de les difficultés d'interprétation que l'on rencontre si l'on essaye de

mener l'analyse directement sur la distribution réelle des $\{n_{i,j}\}$, comme dans le paragraphe 5.1.

Figure 9 : évolution du ratio (réel/benchmark) pour les b_i (à gauche) et les s_j (à droite)

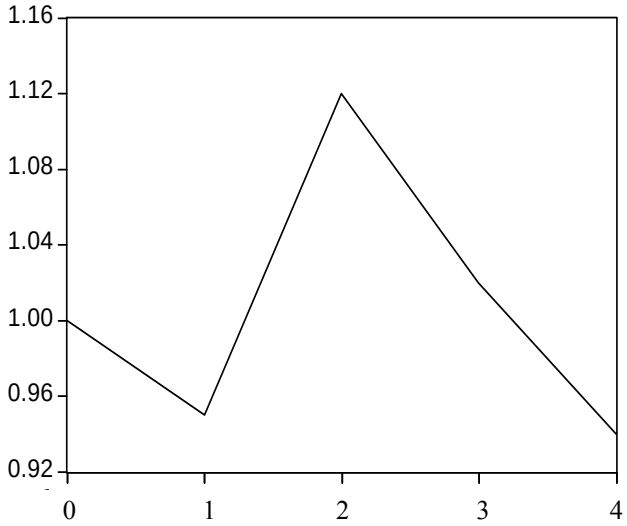


Un niveau supérieur à 1 correspond à une situation où les transactions (achat ou vente selon l'indicateur étudié) ont été supérieures à la moyenne et inversement si le ratio est inférieur à 1. Ainsi, on détecte à la date 2 un niveau d'activité important pour les vendeurs, probablement en raison de la forte hausse des prix immobiliers (h_2 = 1.284). Les investisseurs souhaitent prendre leurs bénéfices. Dans le même temps, l'indicateur d'achat est sous le niveau moyen. Les prix étant très élevés, les acheteurs ont pu trouver un intérêt financier à retarder leur transaction en espérant de meilleurs jours. A la date t = 3, la situation s'inverse complètement (pour plus de précisions sur ces indicateurs on pourra se reporter au chapitre 2, paragraphe 3.3.2).

On peut également étudier le ratio des périodes de détention τ entre le benchmark et l'échantillon réel pour détecter d'éventuels accroissements ou d'éventuelles diminutions de la durée de détention, cf figure 10. Le phénomène le plus évident dans ce graphique est le pic observé à la date t = 2 pour les investisseurs qui détenaient de l'immobilier au cours de l'intervalle [2,3]. Cette population a subi

de plein fouet l'effondrement des prix de la date 3 et il est probable qu'à la suite de ces événements les candidats à une revente rapide ont retardé substantiellement leur transaction. Pour τ_1 la situation est inversée[45].

Figure 10 : évolution du ratio τ-réel / τ-benchmark



La situation étudiée ici est bien sûr une caricature, mais la méthode développée au cours de ce paragraphe semble prometteuse car elle fournit des indicateurs pour les niveaux d'activité du marché, les incitations à l'achat et à la revente et les variations de la période de détention. On retrouve par l'analyse des seules données les comportements des acteurs du marché, comme présentés dans le paragraphe 5.1 d'une façon omnisciente. Cette méthode permet donc de remonter des données (observables) aux causes premières (inobservables au premier abord) d'une manière relativement fiable.

[45] La faible valeur observée pour τ_4 est sans soute plus une conséquence d'un effet de bord qu'un véritable phénomène économique.

6. Un nouveau formalisme pour l'indice de ventes répétées : synthèse

On adopte dans ce paragraphe une démarche formaliste et non plus exploratoire. Le but de cette partie est de présenter synthétiquement et en toute généralité, la nouvelle formulation du RSI qui s'est dégagée au cours des paragraphes précédents. Les concepts nécessaires à cet exposé sont très similaires à ceux du cas simplifié, la structure fondamentale étant quasiment la même. Le lien entre R et H apparaîtra maintenant comme un corollaire et non plus comme la problématique principale.

6.1. La levée des simplifications et le temps d'égalité des bruits

Jusqu'à présent deux simplifications ont été faites pour étudier le problème de minimisation, d'inconnues $R = (r_0, \ldots, r_{T-1})$:

$$\mathrm{Min}_R \left[\sum_{i<j} \sum_{k'} \{\ln(p_{k',j} / p_{k',i}) - (r_i + \ldots + r_{j-1})\}^2 / \{ \sigma_G^2 (j - i) + 2\sigma_N^2 \} \right] \quad (7'')$$

On a admis premièrement que toutes les transactions étaient réalisées au niveau de l'indice de prix H : $p_{k',j} = h_j$ et $p_{k',i} = h_i$. Cette hypothèse de travail a très nettement simplifié le problème de minimisation car dans une telle situation tous les rendements des ventes répétées de la classe (i,j) deviennent égaux à $\ln(h_j/h_i)$. Comme il sera démontré en annexe (annexes 6 à 10), renoncer à cette hypothèse amènera à remplacer les valeurs h_i et h_j par les moyennes géométriques des prix d'achat et des prix de revente pour les transactions de la classe (i,j) :

$$h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}} \qquad h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}} \qquad (33)$$

Deuxièmement, nous avons aussi supposé que le bruit représentant les imperfections du marché était nul : $\sigma_N = 0$. Il faut maintenant prendre en compte le bruit dans son ensemble, $2\sigma_N^2 + \sigma_G^2 (j - i)$. Les mesures d'information seront par conséquent modifiées. Pour une meilleure appréhension et un usage plus aisé de cette quantité il est intéressant de mettre σ_G^2 en facteur :

$$2\sigma_N^2 + \sigma_G^2 (j - i) = \sigma_G^2 [(2\sigma_N^2/ \sigma_G^2) + (j - i)] = \sigma_G^2 [\Theta + (j - i)] \quad (34)$$

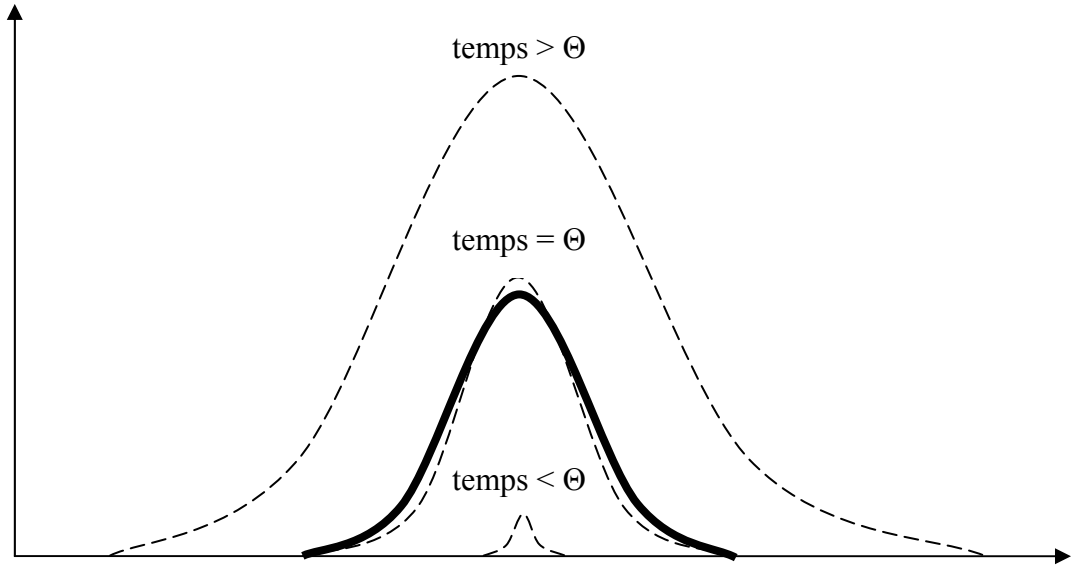
Que représente le paramètre $\Theta = 2\sigma_N^2/ \sigma_G^2$? La première source de bruit modélisée par le bruit blanc $N_{k,t}$ fournit une intensité constante, la seconde associée à la marche aléatoire gaussienne $G_{k,t}$ voit sa contribution varier temporellement. Au début la source constante est la plus importante mais, au fur et à mesure du passage du temps, la part de la seconde augmentant, il arrive un moment où ces deux bruits atteignent le même niveau, cf figure 11. A partir de cette date la source gaussienne devient prépondérante. L'instant précis où la relation s'inverse est solution de l'équation :

$$2\sigma_N^2 = \sigma_G^2 * \text{temps} \Leftrightarrow \text{temps} = 2\sigma_N^2/ \sigma_G^2 = \Theta \quad (35)$$

Le paramètre Θ sera donc appelé le temps d'égalité des bruits.

Comme il a déjà été mentionné, la notion de bruit est une notion relative : ce qui importe ce n'est pas le niveau absolu mais le rapport entre les différents niveaux. En d'autres termes il n'existe pas d'unité de référence pour mesurer le bruit, ou l'information, car ces grandeurs peuvent se définir à un coefficient multiplicatif près. On peut donc décider de travailler directement avec les variances $2\sigma_N^2 + \sigma_G^2 (j - i)$ ou bien, en factorisant par σ_G^2, avec $\Theta + (j- i)$. Cette deuxième possibilité, plus simple, sera privilégiée. Ainsi, pour les $n_{i,j}$ ventes répétées de la cellule (i, j), la quantité d'information associée sera dorénavant $L_{i,j} = n_{i,j} / (\Theta + (j- i))$.

Figure 11 : temps d'égalité des bruits



Trait continu : bruit constant $2\sigma_N^2$ trait discontinu : bruit gaussien σ_G^2 * (temps)

6.2. La décomposition algorithmique de l'indice de ventes répétées

La résolution du problème d'optimisation, en toute généralité, est présentée dans les annexes 6 à 10. Elle s'inspire de la démonstration réalisée dans le cas simplifié, en redéveloppant les raisonnements. On aboutit à un algorithme de calcul relativement simple, et dont les différents constituants sont interprétables économiquement, cf. figure 12. La partie gauche et le paragraphe 6.3 sont associés aux concepts informationnels (typiquement : matrice Î) tandis que la partie droite et le paragraphe 6.4 correspondent à des grandeurs mesurant des prix (typiquement : moyenne des taux moyens P) ; l'indice résulte de la combinaison de ces deux aspects.

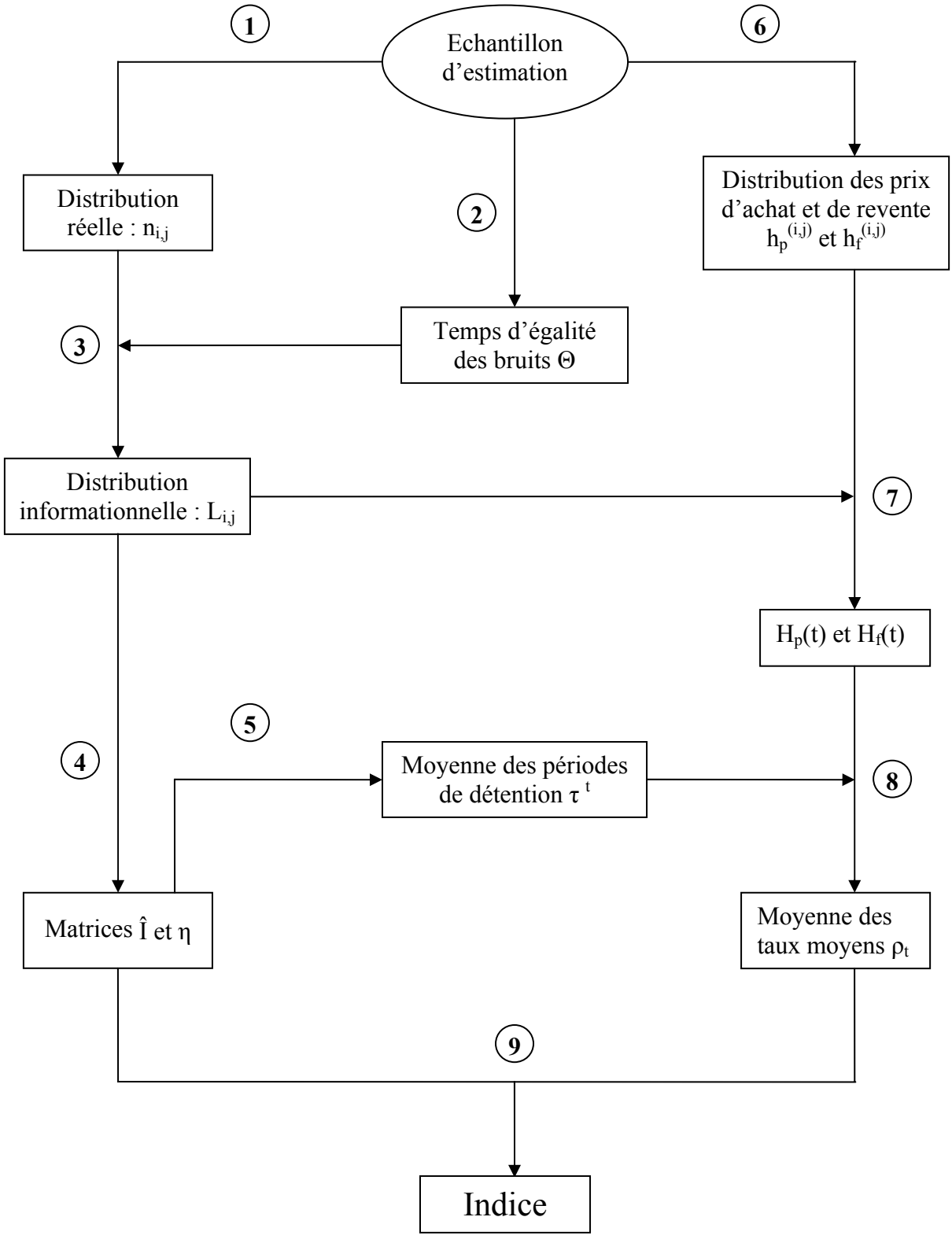
Les paragraphes 6.3 et 6.4 résument les différentes notations nécessaires à la résolution du problème général ; beaucoup d'entre elles seront utiles à l'analyse économique par la suite. Ces notations seront très similaires à celles introduites dans

le cas simplifié et il ne sera donc pas nécessaire de lire en détails la démonstration générale pour saisir le fonctionnement du modèle, l'esprit restant le même.

La structure et l'interprétation des différentes moyennes seront globalement conservées mais on y introduira toutefois des coefficients de pondérations liés à une fonction G (cette fonction rendra compte de l'importance relative des deux sources de bruit ($N_{k,t}$ et $G_{k,t}$) au cours du temps).

L'autre différence notable consistera à remplacer les prix des transactions réalisées au niveau de l'indice, h_i et h_j, par les moyennes géométriques des prix d'achat $h_p^{(i,j)}$ et de revente $h_f^{(i,j)}$. Pour une classe de ventes répétées (i,j) le prix moyen d'achat à "i" passera donc de h_i dans la situation simplifiée à $h_p^{(i,j)}$ dans la situation générale ; et de même pour les reventes.

Figure 12 : Algorithme de calcul de l'indice de ventes



Légende de la Figure 12 (les nouveautés par rapport à la situation simplifiée sont indiquées en caractères gras)

(1) Nombre de ventes répétées dont l'achat est réalisé à t_i et la revente à t_j, organisés dans un tableau triangulaire supérieur

(2) Estimation des volatilités pour le bruit blanc et la marche aléatoire (étapes 1 et 2 de la procédure proposée par Case et Shiller) : σ_N et σ_G. Le temps d'égalité des bruits est défini par : $\Theta = 2\sigma_N^2 / \sigma_G^2$

(3) Distribution informationnelle des ventes répétées obtenue en divisant les $n_{i,j}$ par la période de détention augmentée du temps d'égalité des bruits :

$$\mathbf{L_{i,j} = n_{i,j} / (\Theta + j - i)}$$

(4) La matrice $\hat{I}$ se déduit de la distribution informationnelle en sommant pour chaque intervalle de temps [t,t'] les $L_{i,j}$ pertinents, c'est-à-dire ceux associés à des ventes répétées dont la période de détention inclut [t,t']. Les éléments diagonaux de la matrice diagonale η, se déduisent de ceux de $\hat{I}$ par sommation sur les lignes (ou les colonnes).

(5) On obtient les périodes de détention moyenne τ^t en divisant les éléments diagonaux de $\hat{I}$ par ceux de η.

(6) Pour chaque classe de vente répétée (i,j), on calcule la moyenne géométrique des prix d'achat et des prix de revente :

$$\mathbf{h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}}} \qquad \mathbf{h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}}}$$

(7) Pour l'ensemble des personnes qui détiennent de l'immobilier entre t et t+1, le prix moyen d'achat (resp. le prix moyen de revente) se calcule comme la moyenne géométrique des $h_p^{(i,j)}$ (resp. des $h_f^{(i,j)}$), pondérés par les $L_{i,j}$, pour les classes de ventes répétées pertinentes :

$$\mathbf{H_p(t) = (\Pi_{i \le t < j} (h_p^{(i,j)})^{L_{i,j}})^{1/I^t}} \qquad \mathbf{H_f(t) = (\Pi_{i \le t < j} (h_f^{(i,j)})^{L_{i,j}})^{1/I^t}}$$

(8) La moyenne des rendements moyens réalisés par les propriétaires détenant de l'immobilier entre t et t+1, sur toute la durée de leur investissement, se calcule comme un taux de rentabilité pour un achat au prix de $H_p(t)$, une revente à $H_f(t)$ et une durée de détention τ^t :

$$\mathbf{\rho_t = (1 / \tau^t) * \ln [H_f(t) / H_p(t)] = (I^t / (n^t G(\zeta^t)) * \ln [H_f(t) / H_p(t)]}$$

(9) Les taux de croissance R de l'indice de ventes répétées s'obtiennent grâce à la relation $\hat{I}R = \eta P \Leftrightarrow R = (\hat{I}^{-1} \eta) P$; où P représente le vecteur des ρ_t. **La relation fondamentale est conservée.**

6.3. Résume des notations liées aux distributions réelles et informationnelles

Les différentes notations réunies ici sont associées au coté gauche du schéma précédent

- Les lettres B et b correspondent à “buy at” , S et s à “sell at”
- Les minuscules sont associées à des concepts réels, les majuscules à des concepts informationnels.
- Les définitions sont illustrées avec différentes figures où T = 4
- Spl désigen l'échantillon global des ventes répétées
- La fonction G est définie par $G(x) = x / (x + \Theta)$
  Pour une période de détention de j – i, on a $G (j – i) = (j – i) / (\Theta + (j - i))$.
  Cette quantité peut se comprendre comme la proportion du bruit provenant de la marche aléatoire gaussienne $G_{k,t}$, à savoir (j – i), dans le bruit total $\Theta + (j – i)$.
- Différentes notions de moyennes seront utilisées dans les formules ci-dessous ; on pourra se reporter à l’annexe 13 pour un rappel sur ces concepts.

6.3.1. Les grandeurs de l'échantillon Spl pris dans son ensemble

Tableau 15 : Distributions réelles et informationnelles

$n_{i,j}$	0	1	2	3	4	Somme
0		$n_{0,1}$	$n_{0,2}$	$n_{0,3}$	$n_{0,4}$	B_0
1			$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	B_1
2				$n_{2,3}$	$n_{2,4}$	B_2
3					$n_{3,4}$	B_3
4						
Somme		s_1	s_2	s_3	s_4	N

$L_{i,j}$	0	1	2	3	4	Somme
0		$L_{0,1}$	$L_{0,2}$	$L_{0,3}$	$L_{0,4}$	B_0
1			$L_{1,2}$	$L_{1,3}$	$L_{1,4}$	B_1
2				$L_{2,3}$	$L_{2,4}$	B_2
3					$L_{3,4}$	B_3
4						
Somme		S_1	S_2	S_3	S_4	I

- $n_{i,j}$: nombre de ventes répétées avec un achat à $t = t_i$ et une revente à $t = t_j$
- $N = \sum_{i<j} n_{i,j}$: nombre total de ventes répétées dans l'échantillon
- $L_{i,j} = n_{i,j} / (\Theta + (j - i))$: équivalent informationnel de $n_{i,j}$

  Le dénominateur diffère de celui du cas simplifié par la présence du paramètre Θ qui capture l'intensité du bruit en provenance de la source constante $N_{k,t}$ (le bruit blanc)
- $I = \sum_{i<j} L_{i,j}$: quantité d'information totale, équivalent informationnel de N

- $b_i = n_{i,i+1} + n_{i,i+2} + \ldots + n_{i,T}$: nombre de ventes répétées avec un achat à $t = t_i$ (à ne pas confondre avec b^t défini ci-dessous)
- $s_j = n_{0,j} + n_{1,j} + \ldots + n_{j-1,j}$: nombre de ventes répétées avec une revente à $t = t_j$ (à ne pas confondre avec s^t défini ci-dessous)
- $B_i = L_{i,i+1} + L_{i,i+2} + \ldots + L_{i,T}$: quantité d'information fournit par les ventes répétées avec un achat à $t = t_i$, équivalent de b_i
- $S_j = L_{0,j} + L_{1,j} + \ldots + L_{j-1,j}$: quantité d'information fournit pas les ventes répétées avec une revente à $t = t_j$, équivalent de s_j

- $\sum_{i<j} \sum_{k'} G(j - i) = N\, G(\zeta)$ (36)

  ζ est la G-moyenne des périodes de détention dans l'échantillon global, elle est définie par la relation indiquée ci-dessus.
- $F = I / (N\, G(\zeta)) = (N\, G(\zeta))^{-1} \sum_{i<j} \sum_{k'} G(j - i) * (1 / (j - i))$ (37)

  Moyenne arithmétique des fréquences de détention $1 / (j - i)$ pour les couples de Spl, pondérées par les $G(j - i)$
- $\tau = (I / N\, G(\zeta))^{-1} = F^{-1}$ (38)

  Moyenne harmonique[46] des périodes de détention $(j - i)$ dans Spl, pondérées par les $G (j - i)$.

[46] $(N\, G(\zeta)) / \tau = \sum_{i<j} \sum_{k'} G(j - i) * (1 / (j - i))$

Relations :

Décomposition de N et I, côté vente et côté achat :

$$\mathbf{b} = b_0 + \ldots + b_{T-1} = \mathbf{N} = s_1 + \ldots + s_T = \mathbf{s} \qquad (39)$$

$$\mathbf{B} = B_0 + \ldots + B_{T-1} = \mathbf{I} = S_1 + \ldots + S_T = \mathbf{S} \qquad (40)$$

Lien entre les deux concepts de détention moyenne : $\boldsymbol{\tau} = \boldsymbol{\zeta}$ (41)

6.3.2. Les grandeurs des échantillons $Spl^{[t', t+1]}$ et Spl^{t}

Ce paragraphe concerne l'intervalle [t', t +1], il est organisé de la même manière que le 6.3.1. Les couples de dates (i ; j) pertinents pour cet intervalle sont ceux dont la période de détention inclut [t', t +1], c'est-à-dire tels que $0 \leq i \leq t' \leq t < j \leq T$.

On notera par $Spl^{[t',t+1]}$ le sous-échantillon de Spl constitué de ces ventes répétées pertinentes. Quand t' = t, c'est-à-dire lorsque l'on s'intéressera au $t+1^{eme}$ intervalle de temps élémentaire [t;t+1], certaines notations seront allégées :

$$Spl^{t} = Spl^{[t, t+1]} \qquad n^{t} = n^{[t, t+1]} \qquad I^{t} = I^{[t, t+1]}$$

$$F^{t} = F^{[t, t+1]} \qquad \tau^{t} = \tau^{[t, t+1]}$$

$$b^{t} = b^{[t, t+1]} \qquad s^{t} = s^{[t, t+1]} \qquad B^{t} = B^{[t, t+1]} \qquad S^{t} = S^{[t, t+1]}$$

(b^{t} et s^{t} sont différents des grandeurs b_i et s_j introduits précédemment)

Les tableaux ci-dessous illustrent les concepts suivants pour l'intervalle [1 ; 3], c'est-à-dire pour t' = 1 et t = 2

Tableau 16 : effectifs pertinents et information pertinente pour un intervalle

$n_{i,j}$	0	1	2	3	4	Somme	$L_{i,j}$	0	1	2	3	4	Somme
0		$n_{0,1}$	$n_{0,2}$	$n_{0,3}$	$n_{0,4}$	b^2_0	0		$L_{0,1}$	$L_{0,2}$	$L_{0,3}$	$L_{0,4}$	B^2_0
1			$n_{1,2}$	$n_{1,3}$	$n_{1,4}$	b^2_1	1			$L_{1,2}$	$L_{1,3}$	$L_{1,4}$	B^2_1
2				$n_{2,3}$	$n_{2,4}$	⋮	2				$L_{2,3}$	$L_{2,4}$	⋮
3					$n_{3,4}$	⋮	3					$L_{3,4}$	⋮
4						⋮	4						⋮
			Somme	s^1_3	s^1_4	$n^{[1;3]}$				Somme	S^1_3	S^1_4	$I^{[1;3]}$

- $n^{[t', t+1]} = \sum_{i \le t' \le t < j} n_{i,j}$: nombre de couples pertinents pour [t', t+1]
- $I^{[t', t+1]} = \sum_{i \le t' \le t < j} L_{i,j}$: quantité d'information pertinente pour [t', t+1], équivalent de $n^{[t', t+1]}$

- $b^t_i = n_{i, t+1} + n_{i, t+2} + \ldots + n_{i, T}$: nombre de ventes répétées avec un achat à t_i et une revente à $t_j > t$
- $s^{t'}_j = n_{0, j} + n_{1, j} + \ldots + n_{t', j}$: nombre de ventes répétées avec un achat à $t_i \le t'$ et une revente à t_j .
- $B^t_i = L_{i,t+1} + L_{i,t+2} + \ldots + L_{i,T}$: équivalent informationnel de b^t_i
- $S^{t'}_j = L_{0,j} + L_{1,j} + \ldots + L_{t',j}$: équivalent informationnel de $s^{t'}_j$

- $\sum_{i \le t < j} \sum_{k'} G(j - i) = n^t\, G(\zeta^t)$ (36')

  ζ^t est la G-moyenne des périodes de détention dans l'échantillon Spl^t, elle est définie par la relation ci-dessus
- $F^t = I^t / (n^t\, G(\zeta^t)) = (n^t\, G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} G(j - i) * (1 / (j - i))$ (37')

  Moyenne arithmétique des fréquences de détention 1 / (j – i), pour les couples de Spl^t, pondérées par les G (j – i).

- $\tau^t = (n^t G(\zeta^t)) / I^t = (F^t)^{-1}$ (38')

  Moyenne harmonique[47] des périodes de détention (j – i), pour les couples de Spl^t, pondérées par les G (j – i).

Relations

Décomposition de $n^{[t', t+1]}$ et $I^{[t', t+1]}$, côté vente et côté achat :

$$\mathbf{b^{[t',t+1]}} = b^t_0 + b^t_1 + \ldots + b^t_{t'} = \mathbf{n^{[t',t+1]}} = s^{t'}_{t+1} + \ldots + s^{t'}_{T-1} + s^{t'}_{T} = \mathbf{s^{[t',t+1]}} \quad (39')$$

$$\mathbf{B^{[t',t+1]}} = B^t_0 + B^t_1 + \ldots + B^t_{t'} = \mathbf{I^{[t',t+1]}} = S^{t'}_{t+1} + \ldots + S^{t'}_{T-1} + S^{t'}_{T} = \mathbf{S^{[t',t+1]}} \quad (40')$$

Lien entre les deux concepts de détention moyenne : $\boldsymbol{\tau^t = \zeta^t}$ (41')

6.3.3. *Les matrices*

- η = diag ($n^0 G(\zeta^0)$, $n^1 G(\zeta^1)$, … , $n^{T-1} G(\zeta^{T-1})$) : matrice diagonale des $n^t G(\zeta^t)$

- $\hat{I}$: matrice informationnelle. L'élément (p, q) de la matrice est, $I^{[p-1, q]}$ pour $p \leq q$ et $I^{[q-1, p]}$ pour $p > q$.

- J : matrice de dispersion. L'élément (p, q) de la matrice est $I^{[p-1, q]} / I^{p-1}$ pour $p \leq q$ et $I^{[q-1, p]} / I^{p-1}$ pour $p > q$.

[47] $(n^t G(\zeta^t)) / \tau^t = \sum_{i \leq t < j} \sum_{k'} G(j - i) * (1 / (j - i))$

6.4. Résumé des notations liées aux prix

Les différentes notations répertoriées ici sont associées au coté droit de la figure 12. Un prix de transaction est noté par $p_{k,j}$ ou $p_{k,i}$, la lettre "i" indique une date achat et la lettre "j" une date de revente.

- $r_k^{(i,j)} = \ln(p_{k,j} / p_{k,i}) / (j - i)$ (42)

  Taux moyen continu réalisé pour la k^{eme} vente répétée de la classe (i,j) (achat à "i" , revente à "j")

- $h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}}$ (43)

  Moyenne géométrique des prix d'achat dans la classe (i,j)

- $h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}}$ (44)

  Moyenne géométrique des prix de revente dans la classe (i,j)

- $H_p(t) = \left(\Pi_{i \le t < j} (h_p^{(i,j)})^{L_{i,j}} \right)^{1/I^t} = \left(\Pi_{i \le t < j} (\Pi_{k'} p_{k',i})^{1/(\Theta + (j-i))} \right)^{1/I^t}$ (45)

  Moyenne géométrique des $h_p^{(i,j)}$ pertinents pour [t,t+1], pondérés par les $L_{i,j}$ ou, de manière équivalente, moyenne géométrique des prix d'achat pour les investisseurs détenant de l'immobilier pendant [t,t+1], pondérés par les contributions informationnelles correspondantes $1 / (\Theta + (j - i))$.

- $H_f(t) = \left(\Pi_{i \le t < j} (h_f^{(i,j)})^{L_{i,j}} \right)^{1/I^t} = \left(\Pi_{i \le t < j} (\Pi_{k'} p_{k',j})^{1/(\Theta + (j - i))} \right)^{1/I^t}$ (46)

  Moyenne géométrique des $h_f^{(i,j)}$ pertinents pour [t,t+1], pondérés par les $L_{i,j}$ ou, de manière équivalente, moyenne géométrique des prix de revente pour

les investisseurs détenant de l'immobilier pendant [t,t+1], pondérés par les contributions informationnelles correspondantes $1 / (\Theta + (j - i))$.

- $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \leq t < j} \sum_{k'} G(j - i)\, r_{k'}^{(i,j)} = (1/\tau^t) * \ln[H_f(t)/H_p(t)]$ (47)

  Moyenne arithmétique des taux moyens $r_{k'}^{(i,j)}$ pondérés par les $G(j - i)$. ρ_t représente toujours une rentabilité moyenne pour la population des investisseurs possédant de l'immobilier pendant [t, t+1]

6.5. Commentaires

La construction de l'indice de ventes répétées se fait grâce à la régression $Y = (DA) R^* + \varepsilon$, en supposant que la matrice de variance-covariance des résidus $\sum$ est hétéroscédastique. La solution de ce problème classique est bien connue et peut s'obtenir directement sous une forme matricielle :

$$R = [(DA)' \sum^{-1} (DA)]^{-1} (DA)' \sum^{-1} Y \qquad (48)$$

Ou de manière équivalente grâce aux équations normales :

$$[(DA)' \sum^{-1} (DA)] R = (DA)' \sum^{-1} Y \qquad (48')$$

Cette formulation mathématique du problème peut bien sûr s'interpréter en termes de projection orthogonale mais sa lecture financière n'est pas évidente ; le produit matriciel $[(DA)' \sum^{-1} (DA)]$ n'est par exemple pas un objet économique des plus limpides. Si on l'examine plus en détails on peut toutefois s'apercevoir qu'il correspond exactement à la matrice informationnelle Î définie précédemment. D'autre part le second terme, $(DA)' \sum^{-1} Y$, n'est rien d'autre que le produit matriciel

η P. La relation centrale, $\hat{I}$ R = η P, n'est donc qu'une réécriture des équations normales de la méthode des moindres carrés pondérés.

La réflexion développée ci-dessus n'avait pas pour but d'introduire un nouveau type d'indice ou une nouvelle méthodologie. L'indice de Case-Shiller a été considéré en tant qu'objet d'étude, il est donc normal et rassurant de retrouver les mêmes équations. Mais on pourrait alors légitimement se demander si ces efforts de calculs n'ont pas eu pour seul résultat de redémontrer quelque chose de déjà connu ? Est-ce que la relation $\hat{I}$ R = η P n'est pas du même niveau que la découverte de l'eau tiède ? En fait, le principal avantage de l'écriture $\hat{I}$ R = η P est son interprétabilité : la matrice $\hat{I}$ est une matrice d'information, η dénombre les effectifs pertinents[48] pour un intervalle de temps [t,t+1] et le vecteur P donne les rendements moyens réalisés par la population des propriétaires qui détenaient un bien à une date donnée. Cette formulation a un sens économique, contrairement à la solution traditionnelle du problème des moindres carrés qui reste assez obscure financièrement.

La technique de la régression économétrique a été formalisée de manière suffisamment poussée pour pouvoir s'appliquer à des situations très diverses. Dans le cas des indices immobiliers la théorie générale est très précieuse mais il semble aussi intéressant de pouvoir la particulariser. On peut bien sur prendre les données empiriques, les faire passer dans la boîte noire des moindres carrés et obtenir en sortie les valeurs indicielles, mais on se prive alors d'une part non négligeable de l'information enchâssée dans les données réelles. Ouvrir la boîte et voir quels en sont les rouages et les briques élémentaires est satisfaisant intellectuellement mais surtout utile économiquement. Les différentes grandeurs ($H_p(t)$, $H_f(t)$, τ^t, ρ_t …) fournissent des renseignements intéressants comme on le verra dans le chapitre 2, paragraphe 3.3. Cette approche permet de plus d'étudier la relation générale R = F(H), entre

[48] aux coefficients $G(\zeta^t)$ près

l'indice de ventes répétées R et les indices de prix H. Elle fournira des formules simples, intuitives et cohérentes pour l'étude de la volatilité et de la réversibilité de l'indice (paragraphes 2 et 3 du chapitre 3). On pourra également étudier en détail, grâce à des simulations s'appuyant sur la décomposition en briques élémentaires du RSI (figure 12), la performance de l'indice et la sensibilité des erreurs aux différents paramètres du modèle (paragraphe 4, chapitre 2). Enfin, en jouant sur la définition de la mesure d'information, on pourra réfléchir à la construction d'une théorie plus générale des indices immobiliers (paragraphe 1, chapitre 4). L'écriture $\hat{I}\ R = \eta\ P$ n'est donc pas une reformulation tautologique de ce qui existe déjà ; elle permet de mieux comprendre et de mieux étudier l'indice de ventes répétées.

Nous terminerons ce paragraphe en mentionnant que pour la situation BMN (Bailey, Muth, Nourse (1963)), qui n'est en fait qu'un cas particulier du modèle Case-Shiller dans lequel seules agissent les imperfections du marché ($\sigma_G = 0$), les formules doivent être légèrement adaptées (cf. annexe 11). La structure fondamentale reste cependant la même. On pourra également se reporter au début de l'annexe 23 pour avoir un exemple détaillé et concret de la mise en œuvre de l'algorithme de calcul (dans un contexte BMN).

6.6. La synthèse des formules et des hypothèses pour le benchmark (dans le cas général)

Comme on l'a déjà vu, l'échantillon benchmark est une référence utile pour analyser des lots de données réelles. Cette section adapte les formules simplifiées du paragraphe 4 au cas général. On rappelle que le benchmark repose sur les hypothèses suivantes :

- Le prix des biens ne varie jamais : $h_t = h_0$ pour tous les t
- La quantité globalement échangée sur le marché à toutes les dates est constante, on la notera K

- Les décisions d'achat et de revente sont indépendantes entre individus
- La longueur de la période de détention suit une loi exponentielle de paramètre $\lambda > 0$; λ ne dépend pas du propriétaire considéré

Comme les prix de l'immobilier restent constants, les taux R et P sont bien sûr nuls. L'intérêt du benchmark ne concerne pas les taux en eux-mêmes mais plutôt la répartition des données dans le temps. L'accent est donc mis sur l'étude de la distribution réelle et de la distribution informationnelle. Le détail des calculs peut être trouvé dans l'annexe 12.

6.6.1. La distribution réelle

La seule modification à prendre en compte pour la définition des distributions concerne le dénominateur des $L_{i,j}$. Le (j – i) du paragraphe 4 doit être remplacé par Θ + (j – i). Par conséquent, la distribution réelle des $\{n_{i,j}\}$ et les variables réelles associées sont inchangées.

α est le taux instantané de survie, K' le nombre de disparitions dans un ensemble de K biens pendant une unité de temps, d(k) le taux de disparition après k unités de temps et $\lambda_{lin}(T)$ le taux linéaire équivalent à λ.

On a : $\alpha = e^{-\lambda}$ $K' = K (1 - \alpha)/ \alpha$ $d(k) = 1 – \alpha^{k}$ $\lambda_{lin}(T) = d(T)/ T$

Et pour $\lambda \approx 0^{+}$ on peut écrire : $(1 – \alpha) / \alpha \approx \lambda$ $K' = K \lambda$

Le nombre total des ventes répétées observées est : $N = K T (1 – \pi)$ (49)

et la proportion manquante : $\pi = d(T) * (\alpha / T (1– \alpha)) \approx \lambda_{lin}(T) / \lambda$ (50)

Tableau 17 : Distribution réelle pour l'échantillon benchmark

	0	1	2	…	t	t + 1	…	T	b_i
0		K'α	$K'\alpha^2$		$K'\alpha^t$	$K'\alpha^{t+1}$		$K'\alpha^T$	$K(1-\alpha^T)$
1			K'α		$K'\alpha^{t-1}$	$K'\alpha^t$		$K'\alpha^{T-1}$	$K(1-\alpha^{T-1})$
2					$K'\alpha^{t-2}$	$K'\alpha^{t-1}$		$K'\alpha^{T-2}$	$K(1-\alpha^{T-2})$
¦									
t						K'α		$K'\alpha^{T-t}$	$K(1-\alpha^{T-t})$
t + 1								$K'\alpha^{T-t-1}$	$K(1-\alpha^{T-t-1})$
¦									
T – 1								K'α	K (1 – α)
T									
s_j		K(1–α)	$K(1-\alpha^2)$		$K(1-\alpha^t)$	$K(1-\alpha^{t+1})$		$K(1-\alpha^T)$	N

Les grandeurs temporellement variables sont toujours :

$$n^t = (K / (1 - \alpha)) * d (T - t) d(t + 1) \qquad (51)$$

$$n^{[t',t+1]} = (K /(1 - \alpha)) * (1 - d (t - t')) d(T- t) d(t' + 1) \qquad (52)$$

6.6.2. La distribution informationnelle

A partir des $\{n_{i,j}\}$ on obtient la distribution informationnelle en divisant le nombre de ventes répétées $n_{i,j}$ par la mesure de bruit correspondante $\Theta + (j - i)$, cf. tableau 18.

On notera :

$$u_n = \alpha /(\Theta + 1) + \alpha^2 /(\Theta + 2) + \alpha^3 /(\Theta + 3) + \ldots + \alpha^n /(\Theta + n) \qquad (53)$$

$$\ell = \text{Lim } u_n = - \left[\ln (1 - \alpha) + \alpha + \alpha^2 / 2 + \ldots + \alpha^{\Theta} / \Theta \right] / \alpha^{\Theta} \qquad (54)$$

Tableau 18 : Distribution des $\{L_{i,j}\}$ pour le benchmark

	0	1	2	…	t	t + 1	…	T	B_i
0		$K'\alpha/(\Theta+1)$	$K'\alpha^2/(\Theta+2)$		$K'\alpha^t/(\Theta+t)$	$K'\alpha^{t+1}/(\Theta+t+1)$		$K'\alpha^T/(\Theta+T)$	$K' u_T$
1			$K'\alpha/(\Theta+1)$		$K'\alpha^{t-1}/(\Theta+t-1)$	$K'\alpha^t/(\Theta+t)$		$K'\alpha^{T-1}/(\Theta+T-1)$	$K' u_{T-1}$
2					$K'\alpha^{t-2}/(\Theta+t-2)$	$K'\alpha^{t-1}/(\Theta+t-1)$		$K'\alpha^{T-2}/(\Theta+T-2)$	$K' u_{T-2}$
¦									
t						$K'\alpha/(\Theta+1)$		$K'\alpha^{T-t}/(\Theta+T-t)$	$K' u_{T-t}$
t + 1								$K'\alpha^{T-t-1}/(\Theta+T-t-1)$	$K' u_{T-t-1}$
¦									
T – 1								$K'\alpha/(\Theta+1)$	$K' u_1$
T									
S_j		$K' u_1$	$K' u_2$		$K' u_t$	$K' u_{t+1}$		$K' u_T$	I

La formule (54) qui précise la valeur de ℓ ne s'applique que pour des valeurs entières de Θ. Or en pratique ce n'est pas toujours le cas. Ainsi, dans Case,Shiller (1987) quatre indices urbains sont calculés en utilisant la procédure traditionnelle en trois étapes présentée au paragraphe 2.1. La deuxième fournit des estimateurs pour $2\sigma_N^2$ et σ_G^2 desquels on peut déduire les valeurs des temps d'égalité des bruits $\Theta = 2\sigma_N^2/\sigma_G^2$. Et, comme on le constate dans le tableau 19, elles ne sont pas entières.

Tableau 19 : Temps d'égalité des bruits dans Case, Shiller (1987)

City	Atlanta	Chicago	Dallas	San Francisco
Θ	12.89	9.11	6.77	4.20

Si l'on arrondit Θ à l'entier le plus proche l'erreur n'est pas insignifiante : substituer 4 au 4.20 de San Francisco représente par exemple une erreur de 5%. Comme toutes les variables informationnelles dépendent du paramètre Θ, une telle

imprécision est peu souhaitable. Mais, d'un autre côté, avoir une valeur entière simplifie beaucoup les calculs[49]. Pour tenir compte de cette opposition, on supposera donc que Θ est un entier dans le modèle, mais en pratique on pourra corriger les résultats par une interpolation linéaire.

En termes plus concrets, si les deux entiers les plus proches de Θ sont Θ_{inf} et Θ_{sup}, on peut trouver un réel x de [0;1] tel que $\Theta = x\ \Theta_{inf} + (1 - x)\ \Theta_{sup}$. Si l'on doit calculer la valeur d'une expression dépendant de Θ, $f(\Theta)$, on calculera d'abord $f(\Theta_{inf})$ et $f(\Theta_{sup})$ et on considérera que $f(\Theta) \approx x\ f(\Theta_{inf}) + (1- x)\ f(\Theta_{sup})$. Ainsi $\Theta = 4{,}2$ donne $\Theta_{inf} = 4$, $\Theta_{sup} = 5$ et on peut écrire $\Theta = 0.8\ \Theta_{inf} + 0.2\ \Theta_{sup}$. Si l'on suppose que $\alpha = 1$ on a $u_4(\Theta_{inf}) = 0.6345$, $u_4(\Theta_{sup}) = 0.5456$ et l'approximation de $u_4(\Theta)$ sera alors $0.8\ u_4(\Theta_{inf}) + 0.2\ u_4(\Theta_{sup}) \approx 0.6167$. La vraie valeur de $u_4(\Theta)$ est 0.6144, l'erreur commise ici est donc de 0.3%, ce qui représente un niveau très acceptable.

6.6.3. Les variables informationnelles

La quantité d'information totale est : $$I = K' [(T + \Theta + 1)\ u_T - T\ \pi] \qquad (55)$$

L'échantillon d'estimation est constitué à partir des K propriétés négociées à t = 0, des K négociées à t = 1 , … , des K négociées à t = T – 1. S'il n'y avait pas de barrière à t = T toute l'information finirait par être révélée, elle représenterait un niveau total de K'T ℓ. La proportion d'information manquante[50], c'est-à-dire l'équivalent informationnel de π, est donc :

$$\mu = [\ \ell + \pi - (T + \Theta + 1)\ (u_T / T)] / \ell \qquad (56)$$

[49] L'intérêt d'avoir une valeur entière pour Θ apparaît par exemple quand on souhaite déterminer la limite de la suite (u_n), cf. annexe 12.

[50] Dans le paragraphe 4.3 cette formule a été établie pour ψ, la proportion d'information révélée. On travaille ici avec la proportion manquante $\mu = 1 - \psi$.

Pour I^t et $I^{[t',t+1]}$ les notations suivantes doivent être introduites :

$u_n = \alpha /(\Theta + 1) + \alpha^2 /(\Theta + 2) + \alpha^3 /(\Theta + 3) + \ldots + \alpha^n /(\Theta + n)$ avec $u_0 = u_{-1} = 0$

$U(m,n) = u_n - u_m = \alpha^{m+1} / (\Theta + m + 1) + \ldots + \alpha^n / (\Theta + n)$ pour $m \leq n$

$$\mathcal{U}(t', t, T) = (1+ \Theta/2)\, U(t,T) - [\, (t + \Theta/2)\, U(t-t'-1,t) - (t' - \Theta/2)\, U(t-t'-1,T-t'-1) \,] + (T + \Theta/2)\, U(T-t'-1,T) \quad \text{pour } 0 \leq t' \leq t < T$$

On a alors : $$I^{[t',t+1]} = ((1 - \alpha) / \alpha) * (n^{[t', t+1]} + K\, \mathcal{U}(t', t, T)) \qquad (57)$$

$$I^t = ((1 - \alpha) / \alpha) * (n^t - K + K\, \mathcal{U}(t, t, T)) \qquad (58)$$

Les formules sont globalement préservées par rapport à la situation simplifiée, à l'exception des définitions de u_n et $\mathcal{U}(t',t, T)$, où l'impact de Θ est maintenant explicitement pris en compte.

Et finalement les fréquences moyennes sont :

$$F^t = [((1 - \alpha) /\alpha) * (n^t - K + K\, \mathcal{U}(t, t, T))] / (n^t G(\zeta^t)) \qquad (59)$$

$$F = ((1 - \alpha) /\alpha) * ((1 - \mu) / (1 - \pi)) * (\ell / G(\zeta)) \qquad (60)$$

6.7. La relation fonctionnelle entre les deux indices est-elle préservée ?

Dans le paragraphe 3 il a été démontré que si les transactions se font toutes au niveau d'un indice de prix, le RSI est alors une fonction déterministe de cet indice, par l'intermédiaire de P et de Î. Que devient la relation fonctionnelle lorsque cette hypothèse simplificatrice est abandonnée ? Pour répondre à cette question, on

approfondit dans ce paragraphe l'étude de $H_p(t)$ et $H_f(t)$. Il apparaîtra que ces deux quantités sont très liées aux valeurs d'un certain indice de prix H.

6.7.1. Un indice de prix particulier

Si l'on veut construire un indice de prix à partir de l'échantillon des ventes répétées, il faut sélectionner toutes les transactions réalisées à une date donnée t, quelque soit leur nature (achat ou revente). Ce sous-échantillon, noté E_t, est représenté dans le tableau 20. Il inclut tous les couples dont la revente se fait à t (colonne) et tous les couples avec un achat à t (ligne).
Il n'est pas possible d'utiliser directement ces données car elles peuvent présenter un biais. Supposons par exemple qu'un bien ait été acheté à 0 et revendu à t. Le nouveau propriétaire pourrait parfaitement choisir de le revendre à T – 1. Dans une telle situation, le prix de la transaction réalisée à t serait enregistré dans la cellule correspondant à $n_{0,t}$, mais aussi dans celle correspondant à $n_{t,T-1}$. Or, si l'on souhaite calculer le prix moyen des transactions effectuées à la date t, il n'y a pas de raison de compter deux fois cette valeur. Dans l'indice de ventes répétées, ce problème ne se pose pas car la distinction est faite entre l'achat et la revente. Par contre, pour les indices de prix, comme la seule chose qui importe est le niveau de la transaction (indépendamment de son sens), il faudra éliminer les redondances avant de mener le calcul.

De l'ensemble de prix E_t, en éliminant ces répétitions, on obtient un sous-ensemble F_t. On notera par $q_t(i)$ les prix des transactions et par $inf_t(i)$ leur contribution informationnelle[51]. A partir de l'ensemble F_t on peut calculer un indice de prix dont la valeur à t est la moyenne géométrique des $q_t(i)$ pondérés par les $inf_t(i)$:

[51] $inf_t(i) = (\Theta + \text{longueur de la période de détention})^{-1}$

$$h_t^{Inf_t} = \prod_{F_t} [\, q_t(i) \,]^{inf_t(i)} \qquad \text{où } Inf_t = \sum inf_t(i) \qquad (61)$$

Tableau 20 : Sous-échantillon E_t pour l'indice de prix à la date t

	0	1	2	…	t	t + 1	…	T – 1	T
0									
1									
2									
¦									
t									
t + 1									
¦									
T – 1									
T									

Les différentes valeurs de h_t sont rassemblées dans un vecteur $H = (h_0, \ldots, h_T)$. Les pondérations peuvent sembler exotiques pour construire un indice de prix, mais on verra par la suite qu'elles sont en fait implicites à la méthode des ventes répétées.

6.7.2. La relation entre le RSI et l'indice de prix

Il est maintenant possible d'exprimer $H_p(t)$ et $H_f(t)$ en fonction des valeurs de H à un coefficient multiplicatif près (cf. annexe 14 pour la démonstration) :

$$[\, H_p(t) \,] = [\, \prod_{i = 0,\ldots,t} h_i^{B_i^t} \,]^{1/I^t} \exp(\upsilon^t) \qquad (62)$$

$$[\, H_f(t) \,] = [\, \prod_{j = t+1,\ldots,T} h_j^{S_j^t} \,]^{1/I^t} \exp(\nu^t) \qquad (63)$$

Où : $\nu^t \approx 0$, $\upsilon^t \approx 0$, $E[\nu^t] = 0$ et $E[\upsilon^t] = 0$.

Pour la population des propriétaires qui détenaient de l'immobilier pendant [t,t+1], on retrouve ainsi pratiquement les mêmes formules que dans le paragraphe 3 pour les valeurs moyennes d'achat et les valeurs moyennes de revente, $H_p(t)$ et $H_f(t)$. Ces deux quantités peuvent toujours être vues comme les moyennes géométriques pondérées des valeurs passées, ou futures, d'un indice de prix particulier[52]. En d'autres termes, la situation complexe étudiée dans ce paragraphe est équivalente à celle du paragraphe 3, si et seulement si, on définit l'indice de prix H comme indiqué ci-dessus.

Toutefois, mêmes si les formules sont très proches, on peut signaler deux petites différences avec la situation basique. En premier lieu, l'indice de prix qui apparaît dans (62) ou (63) n'est pas vraiment intuitif puisqu'il s'agit d'une moyenne géométrique avec des pondérations inégales définies par les contributions informationnelles. Comme il a déjà été signalé par Shiller (1991), il n'est pas étonnant de rencontrer une structure géométrique dans le contexte du RSI, l'obtention d'indice arithmétique nécessitant une adaptation de la méthodologie. En ce qui concerne les poids, il faut noter qu'ils décroissent avec la longueur de la période de détention. Ainsi, les biens conservés durant de nombreuses années sont légèrement sous-pondérés dans l'indice de prix H. Mais, ici aussi, il s'agit d'une demi surprise puisque d'après la dynamique des prix posée au début de la modélisation, plus le temps s'écoule plus un bien peut potentiellement dévier de la tendance générale sous l'effet de la marche aléatoire. Les détentions courtes sont donc plus significatives que les détentions prolongées et la sous-pondération est alors un moyen de limiter les conséquences des valeurs extrêmes, qui sont plus susceptibles d'apparaître pour les biens à détention longue. La définition de l'indice de prix H, qui pouvait sembler un peu artificielle à première vue, est donc complètement cohérente avec les hypothèses du modèle Case-Shiller.

[52] Les poids B_i^t et S_j^t qui mesurent l'activité du marché d'un point de vue informationnel sont également les mêmes que dans le paragraphe 3.

En second lieu, les expressions pour $H_p(t)$ et $H_f(t)$ sont ajustées par les quantités $\exp(\nu^t)$ et $\exp(\upsilon^t)$. Elles rendent compte du phénomène de variabilité de la moyenne lorsqu'on la calcule sur un sous-ensemble (cf. annexe 14). Cependant, si les données sont suffisamment nombreuses, ces perturbations seront faibles et les coefficients multiplicatifs seront voisins de 1.

En résumé et pour répondre à la question formulée en début de paragraphe, si l'on définit l'indice de prix H comme indiqué ci-dessus, $H_p(t)$, $H_f(t)$, et donc P, restent des fonctions de H. Par l'intermédiaire des matrices $\hat{I}$ et η, il en est alors de même de R. Il faut bien noter que le terme de fonction est ici un peu abusif. Si dans le paragraphe 3, R et H sont liés par une relation fonctionnelle et déterministe, dans la situation générale ce lien est légèrement perturbé par des phénomènes de fluctuation d'échantillonnage (retranscrits par les quantités $\exp(\nu^t)$ et $\exp(\upsilon^t)$). La relation n'est donc plus purement déterministe. Elle restera toutefois centrée autour de la relation fonctionnelle du paragraphe 3.

7. Conclusion

En étudiant la question du lien fonctionnel entre un indice de prix et un indice de ventes répétées, la structure fine de ce dernier s'est révélée. Elle a conduit à modifier la présentation classique du RSI en s'appuyant sur de nouveaux concepts. Cette approche théorique articule explicitement en une structure unifiée plusieurs notions financières intuitives, interprétables et maniables, dont les évolutions peuvent être observées conjointement : taux moyens et taux instantanés d'un investissement immobilier, longueur de la période de détention, quantité d'information fournie par une transaction, niveaux de liquidité côté achat et côté vente, prix moyen d'achat et de revente pour une population de propriétaires investis en immobilier à une date donnée, etc. La construction d'un échantillon benchmark pour les ventes répétées a

également été exposée et ses perspectives d'emploi illustrées. Ce premier chapitre a posé les fondements théoriques sur lesquels vont s'appuyer les applications ultérieures. Celles-ci confirmeront l'intérêt de l'effort d'abstraction et de modélisation entrepris ici.

Chapitre 2

Méthodologie d'analyse de données et étude de la fiabilité de l'indice

1. Introduction

Ce chapitre vient illustrer la maniabilité des résultats théoriques établis précédemment en les employant pour étudier quelques problématiques empiriques. La première application consistera à présenter en détail, et sur des données réalistes, la méthodologie d'analyse de données issue de la décomposition de l'indice en briques élémentaires et de la comparaison échantillon réel / benchmark exponentiel étalonné. On aura au préalable présenté dans le paragraphe 2 la méthode de génération des données synthétiques qui serviront de support à cette analyse.

Les questions suivantes graviteront toutes autour des notions de fiabilité et d'efficacité de l'indice ; il faudra donc définir rigoureusement ce que l'on entend par "fiable". Pour cela on introduira plusieurs indicateurs destinés à mesurer l'erreur indicielle sous différents angles. Il faut remarquer ici que si l'on souhaite parler d'erreur de mesure cela signifie implicitement que l'on connaît les vraies valeurs, or, lorsque l'on travaille avec des données réelles ce n'est jamais le cas. L'étude de l'efficacité d'une méthode ne peut donc se faire réellement que sur des données synthétiques générées à partir d'une situation de référence, le principe de mesure consistant alors à retrouver la situation de référence (inobservable en pratique) par l'intermédiaire des seules données d'observation. L'efficacité de l'indice sera définie comme la sûreté avec laquelle on remonte des manifestations phénoménales aux causes premières. On réemploiera donc le générateur de données de nombreuses fois dans la suite de ce chapitre pour les études de fiabilité. Nous l'appliquerons en particulier pour examiner la sensibilité du RSI au contexte économique, au temps d'égalité des bruits et à la volatilité des prix (paragraphe 4).

Le dernier paragraphe étudiera l'influence de certaines caractéristiques de la distribution des ventes répétées sur la qualité de l'indice. On examinera ainsi l'impact du nombre de données et l'accent sera mis sur les conséquences des situations de données rares que l'on rencontre souvent en pratique, par exemple pour l'immobilier commercial. On analysera les répercussions de l'hétérogénéité dans l'échantillon d'estimation, on quantifiera les effets parfois surprenants des chocs de liquidité on et

pointera une caractéristique sans doute fondamentale pour la fiabilité, à savoir l'asymétrie des données[1]. Certains de ces phénomènes n'ont jamais été évoqués dans la littérature alors qu'ils ne semblent absolument pas être secondaires, par leur impact sur la qualité de l'indice.

2. Le générateur de données

Travailler sur des données réelles est un exercice fondamental de la pratique scientifique lorsque l'on veut analyser et comprendre des phénomènes économiques. La finalité de la réflexion menée ici étant un peu différente, l'usage d'une base de données réelle n'est pas un impératif absolu. En effet, on ne cherchera pas ici à décrire la réalité mais plutôt à tester les capacités d'une méthodologie particulière à appréhender ses manifestations. Pour ce faire, le travail sur des données synthétiques générées par un processus parfaitement connu est plus adapté que le travail sur des données concrètes car ce choix permet de faire varier facilement les conditions fondamentales de l'économie et de tester la réactivité de l'indice à ces multiples fluctuations. Travailler sur une base de données particulière c'est, par contre, courir le risque d'une analyse partielle de la fiabilité de la méthode car elle n'aurait été testée que dans un contexte qui, bien que réel, n'en serait pas moins spécifique.

Il existe également un autre avantage à cette démarche, celui de l'observabilité parfaite. Même si l'analyse est particulièrement poussée sur un échantillon réel, affirmer que les lois fondamentales sont parfaitement connues semble déraisonnable et bien trop ambitieux. Pour ces échantillons, la quantification des erreurs est alors ambiguë car elle mélange la mesure de la performance du modèle à un problème de connaissance imparfaite de la « juste valeur ». Cet écueil peut être contourné avec les échantillons synthétiques, puisqu'ils seront construits précisément à partir de ces

[1] On parlera d'asymétrie dans un échantillon de ventes répétées quand à une date donnée le nombre d'achats observés est très différent du nombre de reventes observées.

« justes valeurs ». Celles-ci seront bien sûr arbitraires mais elles auront l'avantage d'être parfaitement connues ; il sera alors possible de se concentrer uniquement sur la mesure de performance. Sur le plan des détails informatiques, le générateur a été développé sous Excel VBA ; le but de cette section est d'en exposer les principes.

2.1. La structure temporelle

Un échantillon correspond à une période de 20 années. L'observation des prix est faite semestriellement, produisant une série de 41 dates, notées t_0, t_1, … , t_{40}

2.2. La courbe de référence

Pour générer des données cohérentes il est nécessaire de les construire par rapport à une courbe de référence. Ce « vrai prix de l'immobilier » est inobservable en pratique et difficile à fonder théoriquement, comme nous l'avons déjà mentionné dans l'introduction. Pour les besoins de la simulation, on supposera cependant cette courbe connue[2]. Cet objet ne devra pas être considérée comme une référence économique absolue car il ne s'agit ici que de tester la fiabilité d'une méthode et non pas de définir un nouveau concept. La courbe de référence, que l'on qualifiera aussi de pseudo indice, est définie par rapport à une base 100 à la date t = 0. Ses valeurs sont présentées dans le tableau ci-dessous et illustrées graphiquement.

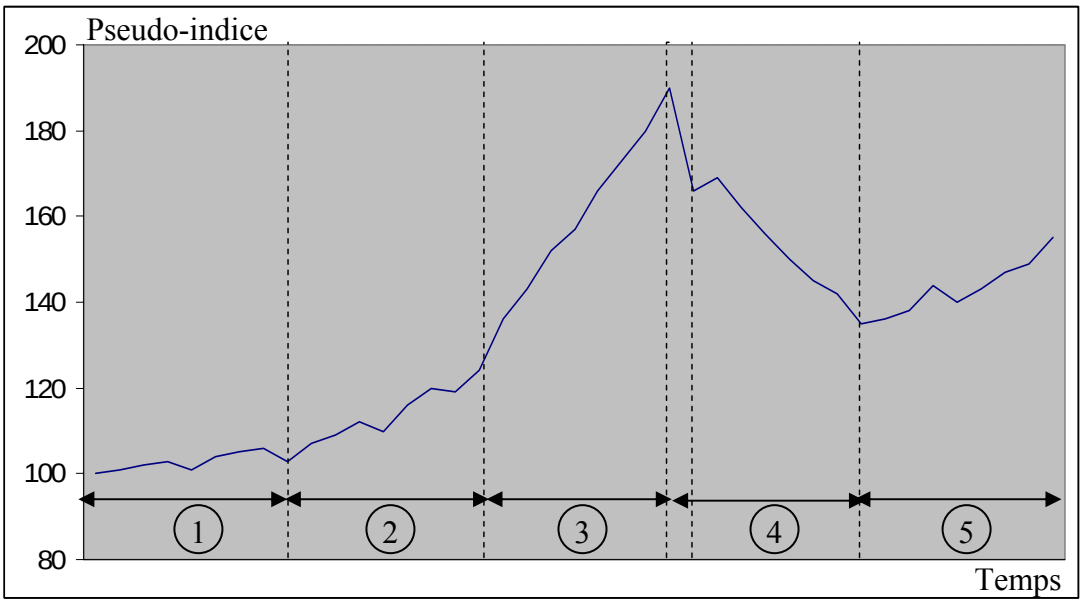
[2] On choisit de le définir ici comme la moyenne des prix de tous les biens existants à une date donnée, en imposant éventuellement une restriction pour avoir une classe d'actifs avec un minimum d'homogénéité. Il faut remarquer que ce choix ne correspond pas à un indice médian. Si le calcul effectué pour ce dernier consiste également en une simple moyenne des prix, elle ne porte que sur les transactions réalisées à une date donnée (flux) et non pas sur l'ensemble de tous les biens immobiliers (stock).

Tableau 1 : valeurs du pseudo indice

Date	valeur	Date	valeur	Date	valeur	Date	valeur	Date	valeur
1	101	9	107	17	136	25	166	33	136
2	102	10	109	18	143	26	169	34	138
3	103	11	112	19	152	27	162	35	144
4	101	12	110	20	157	28	156	36	140
5	104	13	116	21	166	29	150	37	143
6	105	14	120	22	173	30	145	38	147
7	106	15	119	23	180	31	142	39	149
8	103	16	124	24	190	32	135	40	155

Ce tableau décrit un scénario économique fictif mais réaliste composé de cinq périodes de 4 ans que l'on peut retrouver dans l'organisation du tableau et sur le graphique. La première période correspond à un contexte de stagnation des prix de l'immobilier, le taux moyen de croissance par semestre est de 0,38% et il n'atteint que 3% sur l'ensemble des quatre années. Dans la seconde période la croissance des prix s'accélère, mais en restant toutefois à un niveau modéré : 2,37% par semestre en moyenne et 20,39% sur la totalité. La troisième phase se caractérise par une très forte augmentation des prix : taux semestriel moyen de 5,50% et 53,23% sur les quatre années. Le niveau maximal de 190 est atteint à la fin de cette période. Cette date marque le début d'un retournement de tendance, on assiste en six mois à une chute brutale des prix (- 12,63%) qui se poursuit jusqu'à la fin de la seizième année. Le taux moyen par semestre est de - 4,10% , - 28,95% sur toute la période. Le début de la cinquième phase coïncide avec un nouveau retournement de tendance, les prix repartent à la hausse sur un rythme modéré : 1,76% par semestre et 14,81% sur les 4 ans.

Figure 1 : Pseudo indice



Ce scénario reproduit approximativement les mouvements des prix de l'immobilier résidentiel en France entre 1980 et 2005, cf. Baroni et al. (2004). De 1991 à 1997 les prix ont ainsi diminué de 30%, d'après les données réelles. Cette chute est à rapprocher des 28,95% de baisse lors de la quatrième phase du scénario. L'intervalle temporel de l'étude a été ramené ici de 25 ans à 20 ans pour pouvoir faire tourner les modèles sur des échantillons de taille raisonnable.

Dans les paragraphes suivants on pourra être amené à modifier cette courbe en fonction des nécessités de l'étude. Certaines simulations seront ainsi réalisées en supposant que le pseudo-indice est constant et vaut 100 à chaque date.

2.3. Les niveaux d'activité du marché

Pour chaque date de t_0 à t_{39}, on quantifie le nombre total des transactions réalisées sur le marché, relativement au niveau observé à t_0. On définit pour cela un indicateur K_i, pour i variant de 0 à 39, en prenant pour base $K_0 = 1$ (cf. tableau 2 et

figure 2). $K_2 = 1,2$ signifie par exemple que le nombre d'échanges à la deuxième date est supérieur de 20% à celui de la date initiale. En valeur absolue, il suffira donc de connaître le nombre des transactions à t = 0, noté $\mathcal{K}_0$, pour pouvoir en déduire ceux des dates ultérieures par l'intermédiaire de la série des $\{K_i\}$.

Les données de ventes répétées des échantillons d'estimation seront simulées à partir de ces valeurs. Chacune des transactions comptabilisées par les $\{K_i\}$ pouvant être vue comme un achat à la date t_i, deux cas pourront se produire : la revente a lieu avant t_{40} ou après t_{40}. Dans la première situation, on obtiendra une donnée de vente répétée qui servira à estimer l'indice. Par contre, si la vente se produit après t_{40} le bien ne sera pas pris en compte dans l'estimation. Les niveaux K_i donnent l'activité pour le marché dans son ensemble, ils ne correspondent pas directement aux effectifs des échantillons de ventes répétées utilisées pour le calcul de l'indice.

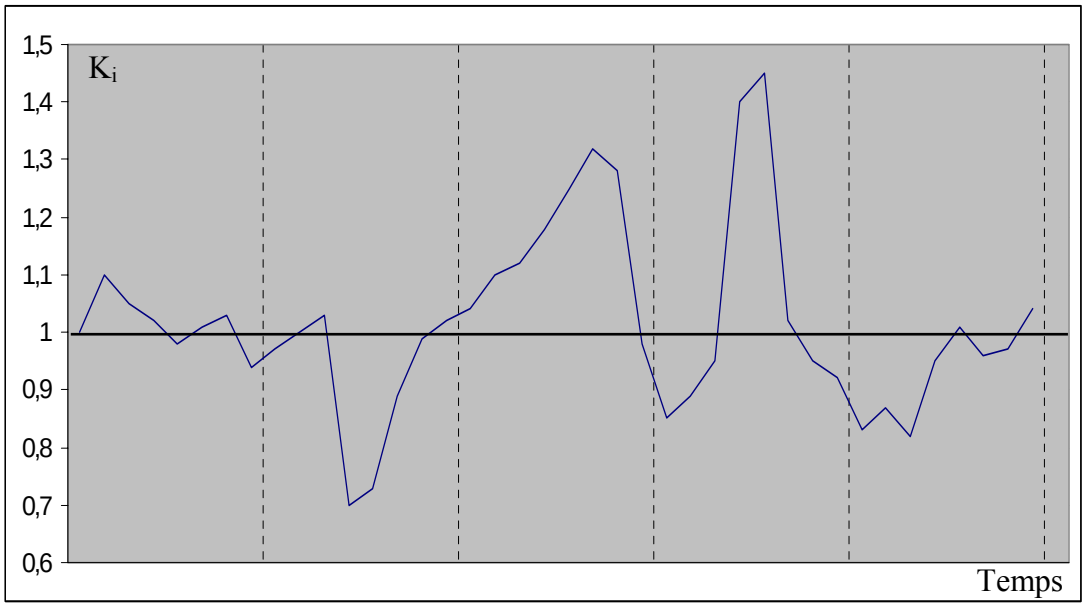
Tableau 2 : niveaux d'activité du marché relativement à t_0

Date	valeur	Date	valeur	Date	valeur	Date	valeur	Date	valeur
0	1	8	0,97	16	1,04	24	0,85	32	0,83
1	1,1	9	1	17	1,1	25	0,89	33	0,87
2	1,05	10	1,03	18	1,12	26	0,95	34	0,82
3	1,02	11	0,7	19	1,18	27	1,4	35	0,95
4	0,98	12	0,73	20	1,25	28	1,45	36	1,01
5	1,01	13	0,89	21	1,32	29	1,02	37	0,96
6	1,03	14	0,99	22	1,28	30	0,95	38	0,97
7	0,94	15	1,02	23	0,98	31	0,92	39	1,04

Pendant les deux premières phases et le début de la troisième, la liquidité du marché est approximativement constante ; la variation par rapport au niveau de référence reste entre +10% et -10%. Deux dates font cependant exception, t_{11} et t_{12}, où l'on constate une forte diminution du nombre de transactions, avoisinant -30%.

Pour donner une interprétation économique à un tel phénomène on pourra par exemple penser à un choc brutal mais temporaire sur les taux longs. Les prêts à taux fixe, majoritaires en France pour l'immobilier résidentiel, étant liés à ces taux de marché, un relèvement brutal des conditions de financement pourrait vraisemblablement amener les particuliers à repousser leur transaction en espérant des jours meilleurs. Le choix de la date d'achat, de la date de revente et donc de la durée de détention du bien permettent aux agents économiques d'agir en fonction de leurs anticipations. S'il y a parfois peu de liberté dans la détermination du prix d'un bien en raison de la pression du marché local, il n'en est pas de même pour le timing.

Figure 2 : Niveaux relatifs d'activité du marché : série des $\{K_i\}$



Le choc sur les conditions de financement ayant été temporaire, on assiste au début de la troisième période à un phénomène de rattrapage qui se traduit par une hausse sensible du volume. Par la suite, le nombre de transactions ne revient pas au niveau usuel, il reste élevé pendant trois ans en fluctuant entre +10% et +32%. Cette période se caractérisera donc par des prix en nette hausse et une activité transactionnelle supérieure à la moyenne. On pourra éventuellement l'analyser

comme la formation d'une bulle. Celle de 1991 s'est par exemple développée dans un contexte semblable, marqué par un nombre de transactions élevé[3].

Peu avant le sommet des prix, la liquidité diminue brutalement : les prix étant jugés trop élevés, le nombre de particuliers prêts à acheter à ce niveau devient plus faible. La chute brutale des prix de -12,63% entre t_{24} et t_{25} va entraîner dans un premier temps un report des transactions. Les particuliers hésitent sur la signification de cet à-coup : simple accident ou tendance durable ? Puis, la baisse des prix se confirmant, un phénomène d'urgence va inciter les vendeurs à agir avant qu'il ne soit trop tard, amenant ainsi les transactions à +40%. A partir de t_{29}, le volume devient faible, traduisant une certaine méfiance vis-à-vis de l'immobilier et un assèchement relatif du marché. A la fin de la cinquième période le niveau normal est à nouveau d'actualité.

Globalement la volatilité du volume est d'abord faible, au milieu du scénario elle augmente significativement, puis elle revient à un niveau standard dans les dernières périodes.

2.4. La modélisation de la décision de revente

Pour chaque achat négocié à t_i il faut modéliser la date de revente. Pour cela on définit pour la population des transactions initiées à t_i la fonction de survie[4] empirique ou, de manière équivalente, le taux instantané de survie d'un semestre à l'autre. Si le contexte était économiquement neutre et déterministe, on pourrait par exemple supposer que 3% du stock encore en vie à l'instant t est revendu au cours du semestre suivant (cela reviendrait à modéliser la durée de détention par une loi exponentielle). Ce niveau de 3% peut être pertinent pour une moyenne de long terme,

[3] Sur le lien volume/prix dans le marché immobilier on pourra consulter Fisher (2003), et plus généralement, Kindleberger (2004) pour les phénomènes de bulles.

[4] La fonction de survie donne à chaque date la proportion des transactions encore en vie ; c'est-à-dire celles pour lesquelles la revente n'est pas encore intervenue.

par contre il ne peut pas s'appliquer directement dans un contexte de court terme sans être modifié par l'environnement économique du moment.

Afin de perturber le modèle déterministe par des phénomènes économiques, on supposera que la date de revente est influencée par le rendement moyen réalisé pendant la détention. Si celui-ci est élevé, l'incitation à la revente est forte ; s'il est faible voire négatif, il constitue un frein (les propriétaires préférant dans ce cas repousser la transaction en espérant pouvoir trouver de meilleures conditions dans quelques mois). Ce rendement moyen ne sera pas calculé au cas par cas pour chaque transaction avec le vrai prix d'achat et le vrai prix de revente. On utilisera simplement un proxy obtenu en calculant la rentabilité grâce au pseudo-indice. Cette approximation n'est pas dommageable car le but poursuivi ici n'est pas de calculer le rendement exact de l'opération mais d'estimer l'ampleur de l'incitation. Une détention entre t_0 et t_{24} donnera par exemple un taux semestriel moyen de 2,67%, entre t_{16} et t_{24} il vaudra 5, 33% et - 4,27% entre t_{24} et t_{32}.

A partir de ces rentabilités on définit les valeurs de la variable d'incitation INC de la façon suivante :

Tableau 3 : Incitations à la revente en fonction de la rentabilité de l'investissement

Indicateur de rentabilité	< -2%	Entre -2% et 0%	Entre 0 et 2%	> 2%
INC	-0,02	-0,01	0	0,01

Enfin, pour éviter un comportement trop déterministe, on introduit un aléa modéré modélisé par une loi de probabilité ε pouvant être uniforme, gaussienne ou log-normale.

Le taux de survie instantané sera donc défini par la formule suivante :

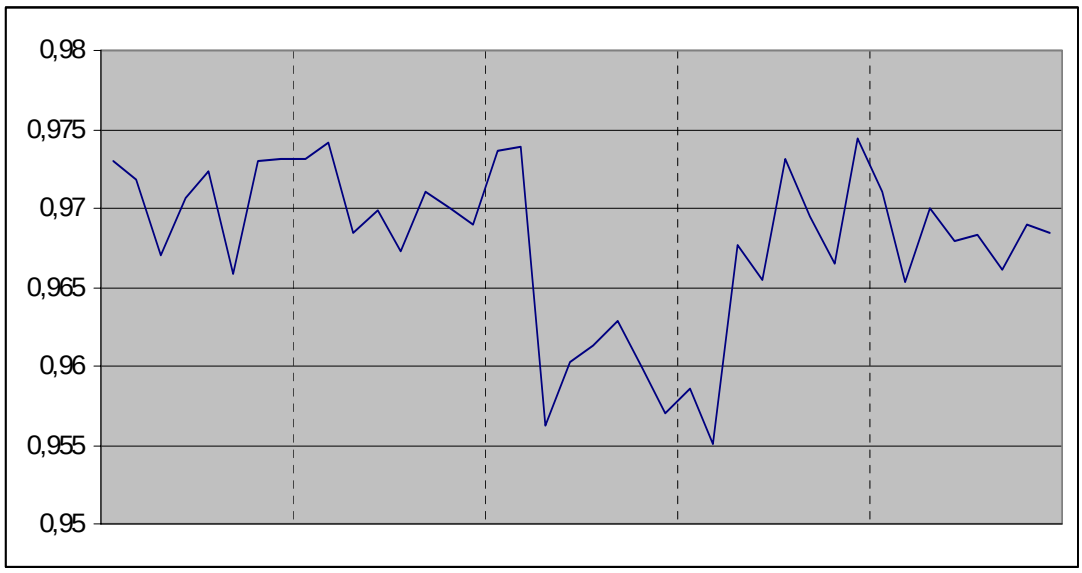
$$\text{Taux de survie entre t et t+1} = 1 - (0{,}03 + INC + \varepsilon) \qquad (1)$$

Pour un effectif Eff_t encore en vie à la date t, l'effectif de la date t+1 se calculera par :

$$Eff_{t+1} = Eff_t * (\text{Taux de survie entre t et t+1}) \qquad (2)$$

Pour les transactions introduites à t_0, 40 taux de survie sont par exemple calculés. Le graphique suivant présente une série de taux de survie possible, obtenue par simulation avec une loi uniforme ε prenant ses valeurs entre -0.005 et +0.005. Cette série semble stationnaire avec une moyenne oscillant autour de 0,97, à l'exception notable de la troisième période et du début de la quatrième où le taux de survie évolue autour de 0,96. L'achat ayant été réalisé à 100, la revente à cette époque où les prix sont au plus haut est particulièrement profitable. Cela se traduit par une valeur de +0,01 pour la variable INC alors qu'aux autres dates elle était restée nulle.

Figure 3 : Exemple de taux de survie pour les transactions initiées à t_0



Si ce procédé de simulation est opérationnel pour les échantillons de grande taille, il posera toutefois un problème pour ceux de taille plus restreinte. Pour simplifier supposons que ε et INC soient nuls, le taux de survie est donc de 97% d'une période à l'autre. Si la cohorte étudiée comporte à la date t un effectif encore en vie de 100 biens, trois vont disparaître entre t et t + 1. Par contre, si l'effectif encore en vie n'est plus que de 10, le résultat du calcul donne 0,3 disparitions. En arrondissant à l'entier le plus proche, cela signifie alors qu'aucun bien n'est revendu. Comme ce raisonnement vaut aussi pour les intervalles ultérieurs, les reventes cesseront donc à partir d'une certaine date sans que l'effectif encore en vie soit épuisé.

La méthode présentée ci-dessus, engendre donc un artefact indésirable pour les petits échantillons : une partie des biens n'est jamais revendue. Si cette caractéristique peut sembler souhaitable et pertinente pour une étude empirique (il n'est en effet pas illégitime de penser que certains biens ne sont jamais revendus, typiquement les propriétés de famille), elle devient toutefois un obstacle pour étudier la sensibilité de l'indice, car elle ne permet pas d'obtenir des lots de très faible taille. En général on utilisera la méthode décrite plus haut pour les simulations des reventes, mais pour l'étude des petits échantillons (paragraphe 5.1) le processus de génération sera modifié comme suit.

A partir d'un effectif de n ventes répétées encore en vie à une date t, on modélise la survie d'un bien jusqu'à la date t + 1 par une variable de Bernouilli de paramètre 0,97. Pour l'ensemble des n biens, en supposant que les décisions sont indépendantes, le nombre de reventes s'obtient alors en simulant une loi binomiale de paramètres 0,03 et n. Cette modélisation permettra de conserver un taux moyen de survie[5] de 0,97 sans engendrer un niveau incompressible sous lequel l'effectif des biens survivants ne pourrait pas descendre.

[5] On ignore ici l'effet de la variable INC

2.5. Les distributions réelle et informationnelle de l'échantillon

Le nombre de transactions réalisées sur le marché à t_0 est fixé arbitrairement à 10000. En générant la fonction de survie empirique comme indiqué précédemment, on peut en déduire le nombre de ventes répétées entre 0 et j ($n_{0,j}$), pour j variant de 1 à 40. Le devenir de la population de biens négociés à t_i, dont le nombre initial s'obtient en multipliant 10000 par K_i, s'obtient par le même procédé. En effectuant cette simulation pour toutes les valeurs de i, on peut ainsi simuler entièrement la distribution des $\{n_{i,j}\}$.

Pour en déduire la distribution des quantités d'information $L_{i,j}$ il suffit de diviser les $n_{i,j}$ par $\Theta + (j - i)$, où Θ est le temps d'égalité des bruits. Afin d'être vraisemblable[6] on fixera cette quantitéà 10. Une fois les $L_{i,j}$ connus, on peut en déduire immédiatement la matrice d'information $\hat{I}$ et la matrice diagonale η.

2.6. Moyenne des taux moyens et RSI

$\hat{I}$ et η ayant été calculées, le dernier élément restant à connaître pour obtenir l'indice de ventes répétées est $P = (\rho_0, \rho_1, \ldots, \rho_{T-1})$, comme indiqué dans l'algorithme de calcul (figure 12, chapitre 1). Pour cela on utilise la relation usuelle :

$$\rho_t = (I^t / (n^t G(\zeta^t))) * \ln [H_f(t) / H_p(t)] \qquad (3)$$

Les quantités I^t et $n^t G(\zeta^t)$ se déduisant directement de $\hat{I}$ et η , $H_f(t)$ et $H_p(t)$ sont donc les seules grandeurs encore inconnues dans cette expression. Celles-ci se calculent grâce aux formules rappelées ci-dessous :

[6] Cf. chapitre 1, paragraphe 6.6.2 pour les valeurs de Θ dans l'article Case, Shiller (1987)

$$H_p(t) = \left(\prod_{i \le t < j} \left(h_p^{(i,j)} \right)^{L_{i,j}} \right)^{1/I^t} \qquad H_f(t) = \left(\prod_{i \le t < j} \left(h_f^{(i,j)} \right)^{L_{i,j}} \right)^{1/I^t}$$

$$\text{où :} \quad h_p^{(i,j)} = \left(\prod_{k'} p_{k',i} \right)^{1/n_{i,j}} \qquad h_f^{(i,j)} = \left(\prod_{k'} p_{k',j} \right)^{1/n_{i,j}} \qquad (4)$$

Dans une classe de ventes répétées (i,j), $h_p^{(i,j)}$ et $h_f^{(i,j)}$ sont les moyennes géométriques équipondérées des prix de revente et des prix d'achat. Dans un souci de simplicité on ne simulera pas tous les prix $p_{k',i}$ et $p_{k',j}$, car en prenant 10000 pour population de base à t_0 l'échantillon total comporte en moyenne 165000 couples de prix[7]. La méthode employée ici consistera à simuler directement les valeurs de ces moyennes à partir du pseudo indice en bruitant ses valeurs aux dates t_i et t_j grâce à un coefficient multiplicatif suivant une loi log- normale $\mathcal{LN}(0 ; 0{,}05)$. Exprimer ainsi $h_p^{(i,j)}$ et $h_f^{(i,j)}$ en unités indicielles et pas en euros ne pose pas de problème particulier comme on peut s'en convaincre en examinant les formules. En effet, le seul élément qui importe dans le calcul de ρ_t est le rendement, c'est-à-dire le rapport, le niveau absolu du prix n'est pas pertinent.

On pourrait penser que dans un échantillon réel les quantités $h_p^{(i,j)}$ et $h_f^{(i,j)}$ étant des moyennes des prix de transaction, elles devraient vraisemblablement être assez proches du « vrai prix » de l'immobilier à la date d'achat t_i et à la date de revente t_j , si l'effectif dans chaque classe (i,j) est suffisamment important. Par contre, si l'effectif devenait plus faible, les phénomènes de fluctuation d'échantillonnage devraient venir amoindrir la qualité d'estimateur de $h_p^{(i,j)}$ et $h_f^{(i,j)}$, la capacité régularisante du grand nombre faisant défaut. En fait, même dans un contexte d'effectif important, les choses ne sont pas aussi catégoriques car il ne s'agit pas du même type de moyenne. Supposons par exemple que les prix soient distribués suivant une loi log-normale de paramètres (μ,σ^2). Le « vrai prix », tel qu'il a été

[7] Cette taille pose des problèmes de temps de calcul s'il faut simuler chaque prix séparément.

défini[8], correspond à la moyenne arithmétique des prix, et vaut donc $\exp(\mu + \sigma^2/2)$ quand le nombre de données est grand. L'espérance de la moyenne géométrique des prix est par contre de $\exp(\mu)$. En d'autres termes $h_f^{(i,j)}$ et $h_p^{(i,j)}$ présentent un biais de $\exp(\sigma^2/2)$. En prenant un niveau standard $\sigma = 0{,}05$ on a $\exp(\sigma^2/2) \approx 1{,}0013$; le biais est donc de – 0,13%. Si la volatilité des prix est raisonnable on pourra donc en pratique ignorer ce décalage, mais si l'échantillon est très dispersé cet élément devra être pris en compte.

Cette caractéristique des indices de ventes répétées a amené Shiller (1991) à les qualifier de géométriques ; on pourra également consulter sur ce sujet Goetzmann (1992) ou Wang, Zorn (1997). Lors de la reformulation théorique développée dans le chapitre 1, différentes grandeurs intermédiaires ont été introduites. On sait dorénavant que celles qui sont à la source[9] de la caractéristique géométrique de l'indice sont $h_p^{(i,j)}$ et $h_f^{(i,j)}$.

Enfin, pour obtenir le résultat final, c'est-à-dire l'indice, il suffira d'appliquer la formule :

$$R = (\hat{I}^{-1} \eta) \; P \qquad (5)$$

En résumé, à partir de la courbe des vrais prix on génère un échantillon ω_0 de ventes répétées permettant de calculer le RSI. La figure 5 du paragraphe 3.3.1 illustre le genre de résultat que l'on obtiendra. La courbe bleue des "vrais prix" n'est pas accessible à l'observateur, sa connaissance du marché se résume à la courbe rouge du RSI, inférée des données.

[8] Cf. note 2 de ce chapitre
[9] $H_f(t)$ et $H_p(t)$ ont bien sur aussi cette structure

3. Une méthodologie d'étude pour les échantillons de ventes répétées

A partir de la méthodologie exposée ci-dessus, un échantillon ω_0 de ventes répétées est généré. Il comporte N = 167999 couples et la quantité d'information totale est I = 9049.1532 , il sera fixé dans toute cette section. Pour mener l'analyse, on étalonnera dans un premier temps le benchmark exponentiel sur ces données. On construira ensuite des indicateurs économiques résumant la situation fondamentale des acheteurs et des vendeurs au cours du temps ; ces indicateurs seront inobservables en pratique car ils supposeront une parfaite compréhension du fonctionnement de l'économie. Puis, dans une troisième partie, la méthodologie d'analyse de données sera mise en œuvre. On cherchera notamment à retrouver les valeurs des indicateurs fondamentaux par l'intermédiaire des seules grandeurs observables.

3.1. L'étalonnage du benchmark exponentiel

L'échantillon exponentiel est entièrement déterminé par deux paramètres : K le niveau d'activité du marché, supposé constant à toute date, et le paramètre de la distribution exponentielle mesurant la vitesse de revente des biens $\alpha = e^{-\lambda}$. On rappelle que $K' = K (1 - \alpha) / \alpha$ et que le taux de survie après k intervalles de temps unitaires est $d(k) = 1 - \alpha^k$. La fonction G mesurant l'importance relative des deux sources de bruit est définie par $G(x) = x / (x + \Theta)$ et donc ici par $G(x) = x / (x + 10)$.

L'étalonnage consiste à choisir le couple (K , α) de telle sorte que les agrégats (N , I) soient identiques à ceux de l'échantillon[10] ω_0. Comme $F = I / (N\ G(\zeta))$ on peut

[10] Comme mentionné précédemment, l'étalonnage n'est pas unique et elle peut être adaptée aux exigences de l'analyse économique. On peut par exemple prendre comme référence le nombre d'achats et de ventes à une date donnée et étalonner le benchmark en conséquence.

raisonner de manière équivalente sur le couple (N , F G(ζ)) , on cherchera donc (K, α) tels que N = 167999 et F G(ζ) = I / N = 9049,1532 / 167999 = 0,053864328.

On sait que pour le benchmark :

$$N = K\,T\,(1-\pi) \quad \text{et} \quad F\,G(\zeta) = \left[(1-\alpha)/\alpha\right] * \left[(1-\mu)/(1-\pi)\right]^{\ell} \qquad (6)$$

π et μ représentent les proportions non révélées, en termes d'effectif et en termes d'information[11] et on a d'après les formules établies dans le chapitre 1 :

$$\pi = d(T) * (\alpha / T(1-\alpha)) \quad \mu = \left[\ \ell + \pi - (T + \Theta + 1) * (u_T / T)\right] / \ell \qquad (7)$$

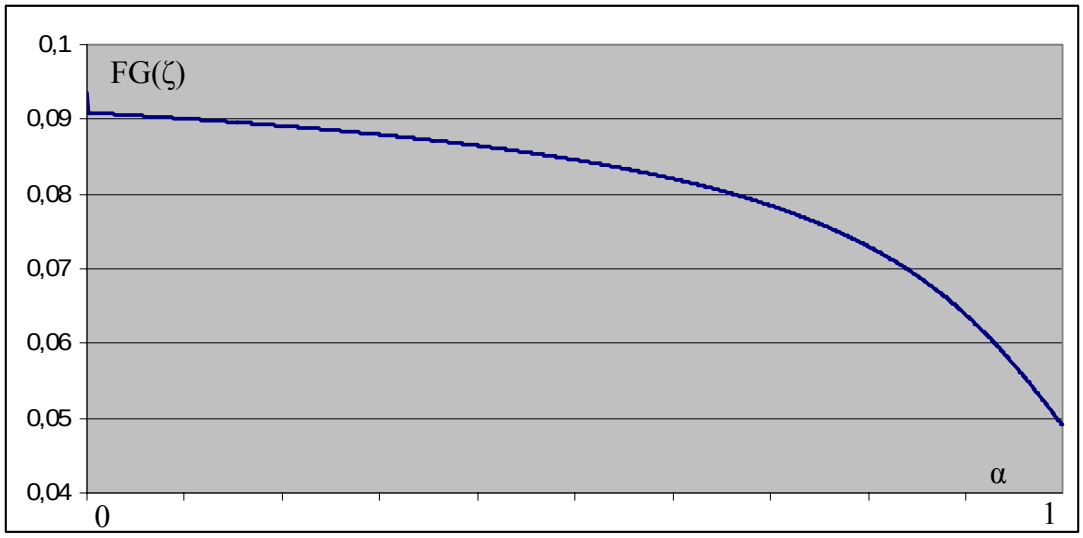
où la suite (u_n) et sa limite $\ell = \text{Lim}\ u_n$ sont définies par :

$$u_T = \alpha/(\Theta+1) + \alpha^2/(\Theta+2) + \ldots + \alpha^T/(\Theta+T)$$
$$\ell = -\left[\ln(1-\alpha) + \alpha + \alpha^2/2 + \ldots + \alpha^{\Theta}/\Theta\right] / \alpha^{\Theta} \qquad (8)$$

La formule donnant l'expression de F G(ζ) ne dépendant que de l'inconnue α, il est possible de s'en servir pour en inférer ce paramètre de manière approchée. La figure 4 présente le graphique que l'on obtient pour cette fonction selon les valeurs de α. La courbe étant strictement décroissante[12], il n'existera donc au plus qu'une solution pour l'équation FG(ζ) = 0,053864328. Il faut remarquer que pour des valeurs inférieures à 0,049 il n'existe pas de α satisfaisant, certains échantillons réels ne peuvent donc pas être associés à des échantillons exponentiels comme on a déjà pu le constater dans le cas simplifié du chapitre 1 (paragraphe 5.2). Le niveau cible F G(ζ) = 0,053864328 est atteint à 10^{-8} près[13] pour α = 0,970577, le taux de disparition instantané correspondant est voisin de 2,94%.

[11] Ces mesures sont associées aux couples dont la date d'achat est entre 0 et T mais avec une date de revente postérieure à T
[12] Et concave
[13] α = 0,970577 donne 0,05386429 et α = 0,970576 donne 0,05386446

Figure 4 : F G(ζ) en fonction de α pour un échantillon benchmark (T = 40, Θ = 10)



Ce résultat est cohérent car, comme les taux de survie de l'échantillon réel ont été simulés grâce à la formule (1) :

$$\text{Taux de survie entre t et t+1} = 1 - (0{,}03 + INC + \varepsilon)$$

le taux de disparition instantané réel moyen devrait donc être voisin de 3%.

Une fois que la vitesse de revente est connue, on peut en déduire immédiatement la valeur de K en utilisant la formule $N = KT(1-\pi) \Leftrightarrow K = N/(T(1-\pi))$; on obtient $K \approx 9880,\ 86$. Cette valeur est à rapprocher de la moyenne des $\{K_i\}$ de l'échantillon réel. Celle-ci étant de 10152.5, l'étalonnage semble donc raisonnable. Une fois ces deux paramètres connus, l'ensemble de la distribution des ventes répétées peut alors être facilement obtenu grâce à la formule $n_{i,j} = K'\alpha^{\,j\,-\,i}$. Les valeurs des agrégats de l'échantillon exponentiel engendré atteignent les niveaux visés de manière très satisfaisante : $N_{exp} = 167999$ et $I_{exp} = 9049{,}1473$.

3.2. La quantification des incitations fondamentales inobservables

La construction d'un échantillon synthétique signifie indirectement que le fonctionnement de l'économie est limpide et que les causes des phénomènes sont parfaitement connues. Par contre dans une situation d'analyse réelle, l'observateur n'étant pas omniscient, il ne peut percevoir que les conséquences de celles-ci. Il doit alors, par le travail économétrique, remonter des manifestations aux causes premières. Les deux indicateurs présentés ci-dessous, $REVENTE_t$ et $ACHAT_t$, sont des grandeurs inobservables résumant la situation fondamentale. Ils serviront de référence pour tester l'efficacité des indicateurs observables que l'on pourra obtenir par la méthode des ventes répétées.

Ces deux indicateurs, $REVENTE_t$ et $ACHAT_t$, renseigneront sur les niveaux des achats ou des reventes à une date donnée. Sont-ils inhabituels, ou bien peuvent-ils être considérés comme normaux ? Pour les définir on considérera que le flux de transactions, à une date donnée, résultera du niveau d'incitation et du stock soumis à cette incitation.

3.2.1. L'incitation à la revente à la date t

La probabilité standard de revente d'un bien au cours de la prochaine période[14] est de 3%. La variable INC permet de faire varier ce niveau en fonction du contexte économique produisant des incitations plus ou moins fortes. Si l'on se place à une date t, l'incitation à la revente à laquelle seront soumis tous les biens achetés à une même date t' < t et encore en vie, est identique par définition de INC. Par contre, pour les biens achetés à une autre date t'', l'incitation de date t sera probablement différente. Pour pouvoir définir la valeur moyenne de l'incitation à la revente à la date t pour l'ensemble des biens, il faudra donc tenir compte des différentes

[14] Taux de revente entre t et t+1 = 0,03 + INC + ε ; 0,03 en moyenne si on ignore l'effet de INC

incitations et des effectifs encore en vie dans chaque cohorte pour pondérer les contributions respectives.

La grandeur, notée $INC_p(t)$, est définie de la manière suivante :

$$INC_p(t) = (\Sigma_{0 \leq k < t} INC(k ; t) K(k ; t - 1)) / (\Sigma_{0 \leq k < t} K(k ; t - 1)) \quad (9)$$

où :

- INC(k ; t) : valeur de la variable INC à la date t pour la cohorte[15] k
- K(k ; t – 1) : effectif de la cohorte k encore en vie après la date[16] t – 1

$INC_p(t)$ exprime une incitation moyenne, elle ne peut pas être interprétée directement en termes d'activité de revente. Imaginons par exemple une situation où le lot de biens soumis à $INC_p(t)$ est de très petite taille, car les reventes ont été nombreuses dans un passé antérieur à t. Le niveau de revente effectif sur le marché sera alors faible, et cela même si l'incitation moyenne est très forte. Pour construire un indicateur du niveau de revente il ne faut donc pas seulement tenir compte de $INC_p(t)$. Le stock susceptible d'être soumis à l'incitation doit aussi être inclus dans le calcul (il le sera par comparaison avec un stock normalisé).

3.2.2. L'indicateur $REVENTE_t$

Le niveau moyen des cohortes K_i à leur date initiale est de 10152,5 (inobservable). Le taux de survie standard d'une période à l'autre est de 0,97 (inobservable). Dans une situation économiquement neutre, le nombre de biens survivants pour la cohorte 0 à la date[17] t^- doit être voisin de $10152,5\ (0,97)^{t-1}$, pour la cohorte 1 de $10152,5\ (0,97)^{t-2}$, pour la cohorte 2 de $10152,5\ (0,97)^{t-3}$… Ainsi

[15] La cohorte k désigne l'ensemble des transactions pour lesquelles l'achat a été réalisé à la date k
[16] C'est l'effectif susceptible d'être soumis à l'incitation de la date t
[17] t^- désigne l'instant juste avant la date t, c'est-à-dire juste avant l'enregistrement des nouvelles transactions.

l'effectif normalisé encore en vie, dans un échantillon de ventes répétées, juste avant les reventes de la date t est :

$$\Sigma_{0 \le k < t}\ 10152{,}5\ (0{,}97)^{t-k-1} = (\ 10152{,}5 / 0{,}03) * (\ 1 - (0{,}97)^{t})$$

L'effectif réel étant de $\Sigma_{0 \le k < t}\ K(\ k\ ;\ t-1)$ on définit alors l'indicateur $REVENTE_t$ par :

$$REVENTE_t = \left[(10152{,}5(1 - (0.97)^{t})/(0{,}03\ \Sigma_{0 \le k < t} K(\ k\ ;\ t\text{-}1)\ )\right]*(1 + INC_p(t)/0{,}03) \quad (10)$$

Le terme entre crochets est supérieur à 1 si l'effectif susceptible d'être revendu à t est plus grand que sa valeur normalisée, inférieur à 1 sinon[18]. Le deuxième élément, $1+ INC_p(t)/0{,}03 = (0{,}03+INC_p(t)) / 0.03$, est le rapport entre le taux réel moyen à t (avec les incitations) et le taux de base ; une valeur supérieure à 1 indique que l'incitation est positive.
En croisant ces deux effets cet indicateur permet alors de détecter, grâce à la connaissance exhaustive de la situation fondamentale, les périodes où les niveaux de reventes seront inhabituels[19].

3.2.3. L'*indicateur* $ACHAT_t$

Le calcul de $REVENTE_t$ s'effectue avec des dates d'achat variant entre 0 et t – 1 et une date de revente fixe. On étudie donc le comportement de différentes cohortes à une même date. Pour construire l'indicateur $ACHAT_t$, dont la fonction est de rendre compte du niveau d'achat dans l'échantillon à la date t, la situation est

[18] Si le stock est faible en raison de nombreuses reventes antérieures, le crochet diminuera le niveau de $REVENTE_t$.
[19] C'est-à-dire sensiblement différent de 1.

inversée : la seule cohorte utile sera celle des biens achetés à t dont on étudiera le devenir pour les dates postérieures. On le définit par :

$$ACHAT_t = (K_t / 10152,5) * (1 + INC_f(t) / 0,03) \qquad (11)$$

où : $INC_f(t) = \Sigma_{t < t' \leq T} INC(t ; t') K(t ; t' - 1) \ / \Sigma_{t < t' \leq T} K(t ; t' - 1)$ (12)

Sa structure est identique à celle utilisée plus haut pour les reventes, elle se compose de deux parties. La première mesure la déviation en termes d'effectifs soumis à l'incitation (niveau réel : K_t , niveau moyen : 10152,5). Dans la seconde partie, $INC_f(t)$ représente la valeur moyenne[20] de la variable INC après t. Cette grandeur mesure donc les incitations futures à la revente entre t + 1 et T, pour les biens négociés à t. Le coefficient multiplicatif $1+INC_f(t)/0,03$ permet de normaliser cette mesure autour de la valeur 1 (inférieur à 1 : incitation négative, supérieur à 1 : incitation positive).

3.2.4. Commentaires

Il peut sembler étrange, à première vue, que la mesure du niveau des achats à une date donnée fasse intervenir les incitations à la revente aux dates postérieures. Cette particularité est en fait une conséquence de la nature conditionnelle des indicateurs $ACHAT_t$ et $REVENTE_t$. Ainsi $REVENTE_t$ ne renseigne pas sur le niveau absolu de revente à t, mais sur le niveau de revente à t, sachant que l'achat a été réalisé entre 0 et t – 1. De même $ACHAT_t$ mesure l'achat à t pour les biens dont la revente interviendra au plus tard à T, il est donc naturel de voir apparaître les incitations futures à la revente dans sa définition.

[20] pondérée par les effectifs

Ce conditionnement est en fait imposé par l'échantillon d'estimation puisque les ventes répétées qui le constituent ont une date initiale supérieure ou égale à 0, et une date de revente[21] inférieure ou égale à T. $ACHAT_t$ et $REVENTE_t$ sont des indicateurs conditionnés par l'observation des ventes répétées. En d'autres termes ils donnent des niveaux de référence pour les échantillons, et non pas dans l'absolu.

Il pourrait être intéressant de généraliser ces indicateurs pour obtenir des niveaux absolus et non plus seulement conditionnels, mais il faudrait pour cela résoudre quelques problèmes techniques. Par exemple, dans le calcul de $INC_p(t)$, dont la définition est rappelée ci-dessous, lever le conditionnement imposerait de laisser varier k de $-\infty$ à t et il faudrait alors s'assurer de la convergence des séries[22].

$$INC_p(t) = (\Sigma_{0 \leq k < t} \; INC(k ; t) \, K(k ; t - 1) \;) / (\Sigma_{0 \leq k < t} K(k ; t - 1))$$

Le but poursuivi dans ce paragraphe consistant simplement à tester l'applicabilité et la fiabilité de la méthodologie d'analyse de données, cette piste ne sera pas développée plus avant.

3.3. Les résultats de l'estimation

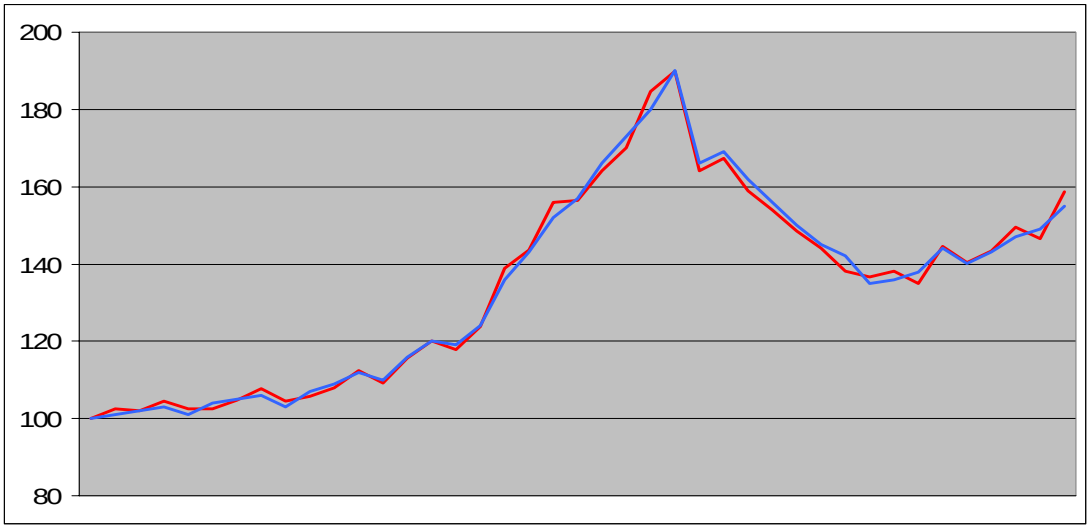
Les valeurs de l'indice et la qualité de l'estimation pour l'échantillon ω_0 sont tout d'abord présentées. Puis, en s'appuyant sur les diverses grandeurs théoriques introduites dans le chapitre 1, on enrichit l'analyse de données. On s'intéressera ainsi aux volumes des transactions, aux prix moyens d'achat et de revente, à la longueur des périodes de détention et aux taux moyens réalisés. Ces divers indicateurs fourniront des informations non négligeables pour étudier le marché.

[21] La génération synthétique de l'échantillon permet de lever certains problèmes d'inobservabilité. Par contre le futur (après T) reste inconnu. On ne peut donc travailler que sur des ventes répétées complètement réalisées (revente avant T) pour mesurer le niveau d'achat.
[22] L'utilisation d'une écriture intégrale pourrait également être intéressante.

3.3.1. Les valeurs indicielles

Le graphique 6 ci-dessous permet de visualiser les résultats obtenus pour l'indice de ventes répétées, estimé sur la base de l'échantillon ω_0 généré par la méthodologie décrite précédemment.

Figure 5 : Indice de ventes répétées pour le scénario benchmark



Courbe bleue : vrai prix Courbe rouge : indice de ventes répétées

Pour cet indice, les indicateurs de fiabilité[23] produisent les valeurs suivantes :

BIAIS(ω_0)	= - 0.01 %	ERREUR_MOYENNE(ω_0)	= 1.16 %
ECART-TYPE(ω_0)	= 0.78 %	ERREUR_MAX(ω_0)	= 2.74 %

Visuellement et numériquement l'indice reproduit donc fidèlement la courbe cible, l'erreur maximale à une date donnée n'étant que 2,74%. Les différentes variables introduites dans la partie théorique vont maintenant permettre d'aller plus loin dans l'analyse et de ne pas se contenter des seules valeurs indicielles.

[23] Ces indicateurs seront définis en détail dans le paragraphe 4.1 ; les noms de ces grandeurs sont toutefois suffisamment explicites pour en comprendre intuitivement la signification

3.3.2. Niveaux d'activité côté achat et côté vente

Les grandeurs $b_i = n_{i,\, i+1} + n_{i,\, i+2} + \ldots + n_{i,\, T}$ (nombre de ventes répétées avec une date d'achat à i et une revente avant T) et $s_j = n_{0,\, j} + n_{1,\, j} + \ldots + n_{j-1,\, j}$ (nombre de ventes répétées avec une date de revente à j et une date d'achat à partir de 0) peuvent être facilement calculées[24] pour l'échantillon ω_0 et pour le benchmark étalonné. Comme ce dernier n'est pas soumis aux fluctuations économiques il fournit des valeurs de référence qui vont être utilisées comme un mètre étalon pour mesurer les activités de revente et d'achat dans l'échantillon réel. On définit pour cela les indicateurs suivants :

$$\begin{aligned} \text{N-revente}_t &= s_t \text{ (échantillon } \omega_0) \,/\, s_t \text{ (échantillon étalonné)} \\ \text{N-achat}_t &= b_t \text{ (échantillon } \omega_0) \,/\, b_t \text{ (échantillon étalonné)} \end{aligned} \qquad (13)$$

En utilisant les équivalents informationnels $B_i = L_{i,i+1} + L_{i,i+2} + \ldots + L_{i,T}$ et $S_j = L_{0,\, j} + L_{1,\, j} + \ldots + L_{j-1,\, j}$ on peut également définir :

$$\begin{aligned} \text{L-revente}_t &= S_t \text{ (échantillon } \omega_0) \,/\, S_t \text{ (échantillon étalonné)} \\ \text{L-achat}_t &= B_t \text{ (échantillon } \omega_0) \,/\, B_t \text{ (échantillon étalonné)} \end{aligned} \qquad (14)$$

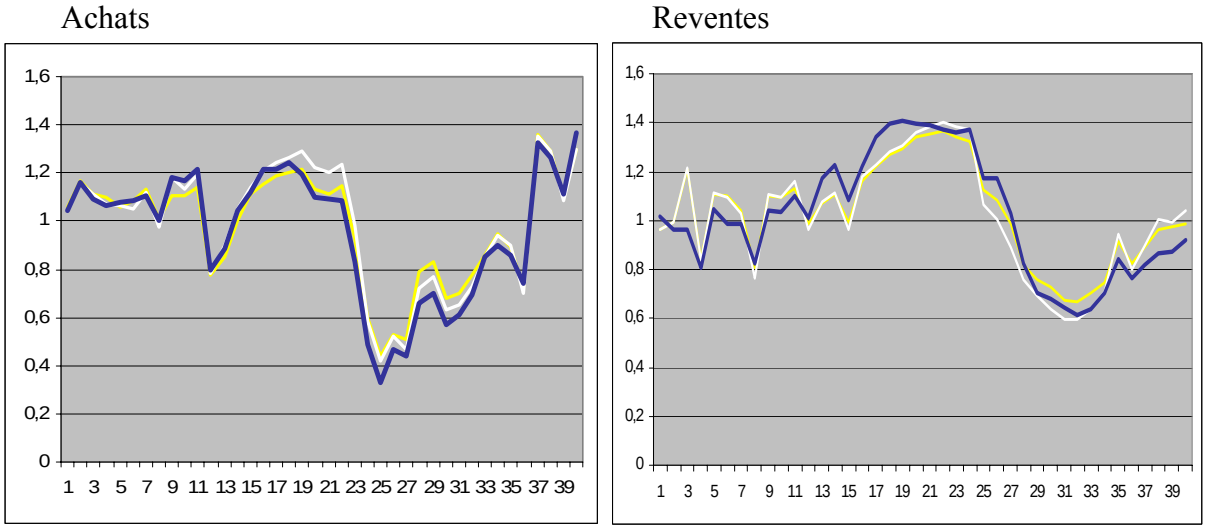
A l'image de REVENTE_t et ACHAT_t, il s'agit d'indicateurs conditionnés par la contrainte imposant d'avoir les deux dates de transaction entre 0 et T. Ils mesurent comme eux les variations des niveaux de revente et d'achat à une date donnée, mais la différence fondamentale entre ces deux types de grandeurs est leur observabilité. REVENTE_t et ACHAT_t reposent sur une connaissance parfaite du fonctionnement de l'économie par l'intermédiaire de la variable INC et des $\{K_i\}$. Ces informations n'étant pas accessible à l'observateur il ne peut donc pas calculer la valeur de ces indicateurs fondamentaux. Par contre le calcul de N-revente_t , N-achat_t , L-revente_t et

[24] Par sommation sur les lignes et les colonnes dans le tableau des $\{n_{i,j}\}$

L-achat$_t$ ne requerrant que les données de ventes répétées et le processus d'étalonnage du benchmark, il est possible de les obtenir facilement. La question naturelle qui se pose alors est de déterminer le degré de fiabilité des indicateurs observables par rapport aux indicateurs fondamentaux.

La figure 6 présente les valeurs obtenues pour ces six grandeurs. Les estimateurs empiriques sont très proches des valeurs théoriques et, à l'exception des reventes de la date 4, les signaux ne produisent pas des informations de sens opposés. On peut calculer pour chacune de ces courbes empiriques l'erreur moyenne en valeur absolue[25]. Pour les achats la N-courbe s'écarte de la courbe théorique de 6,68% en moyenne et la L-courbe de 5,49%, pour les reventes ces chiffres sont respectivement de 6,61% et 7,19%. Si l'on calcule les coefficients de corrélation ils sont de 0,98 et 0,98 pour les achats, 0,95 et 0,93 pour les reventes[26]. Les mesures utilisant les effectifs semblent légèrement supérieures à leurs équivalents informationnels, le gain de précision est cependant faible, voire marginal.

Figure 6 : Indicateurs fondamentaux et indicateurs observables



Bleu : indicateurs théoriques Jaune : N-indicateurs Blanc : L-indicateurs

[25] en utilisant une formule du type ERREUR_MOYENNE(ω), cf. paragraphe 4.1.
[26] Premier chiffre : N-courbe deuxième chiffre : L-courbe

Un niveau moyen d'erreur de 6% pourrait a priori sembler un peu élevé. Toutefois comme ces indicateurs font intervenir des incitations, le jugement doit être nuancé. Un prix ou un rendement sont des objets bien définis sur lesquels il existe un consensus. Cette affirmation ne signifie pas que l'on connaît tous les prix et tous les rendements, car dans ce cas les indices immobiliers n'auraient plus de raison d'être, mais elle cherche seulement à pointer l'absence d'ambiguïtés majeures dans le concept de prix. La notion d'incitation est, par contre, beaucoup plus vague. Une définition précise et consensuelle n'existe probablement pas et le caractère quantifiable de ce concept pourrait même être interrogé, une erreur de 6% n'est donc pas alarmante. Le point central lorsque l'on travaille sur une telle grandeur n'est pas son niveau absolu mais la capture de ses fluctuations. Et sur ce point, les indicateurs empiriques semblent fiables.
Globalement, on pourra donc considérer que la comparaison d'un échantillon ω_0 à son échantillon benchmark étalonné permet de détecter, facilement et de manière fiable, les situations où les volumes de transactions sont inhabituels.

Sur le plan de l'interprétation économique les graphiques de la figure 6 apportent des renseignements intéressants. Ainsi, aux dates t = 11 et t = 12, le scénario comporte un choc sur la liquidité (cf. tableau 2) que les indicateurs d'achat réussissent à capturer (premier creux à 0,8). Par contre les indicateurs de revente ne le détectent pas, probablement en raison du mode de génération des données[27]. Mais, l'évènement majeur du scénario concerne surtout l'envolée des prix jusqu'à la date t = 24, auquel succède une baisse très marquée. Les deux types d'indicateurs réagissent à ce phénomène de manières différentes. Les indicateurs d'achat décroissent nettement dès la date 21, tandis que ceux mesurant la revente ne commencent à diminuer que plus tardivement. On retrouve là les phénomènes qui ont présidé à la génération de l'échantillon : les acheteurs restreignent leur activité avant le sommet des prix, jugeant que ceux-ci ne sont pas soutenables, tandis que les

[27] Le choc, tel qu'il a été modélisé, correspond en effet plutôt à un choc sur l'achat. Pour simuler un choc sur la revente, il aurait fallu réduire les valeurs du taux de revente aux dates t = 11 et t = 12, pour les cohortes antérieures.

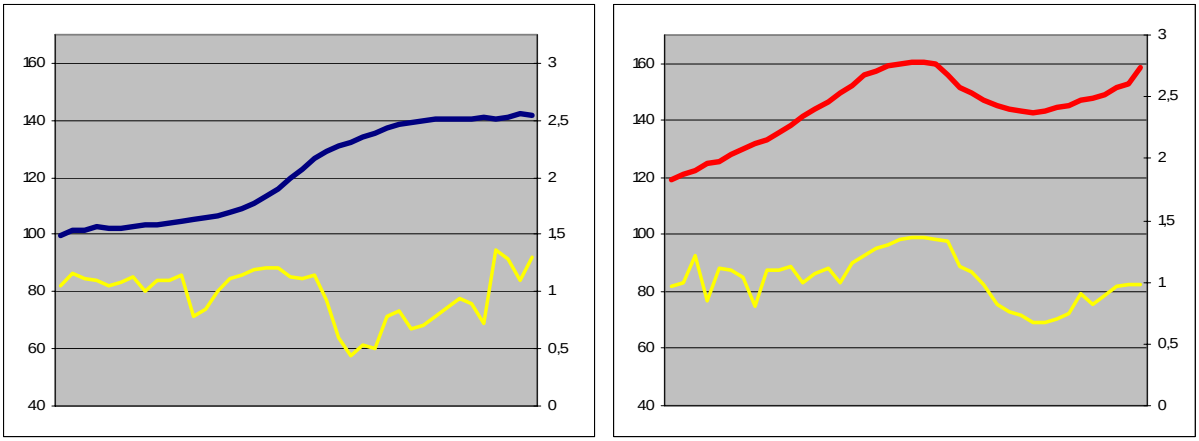
vendeurs retardent leur transaction une fois que les prix se sont nettement effondrés. Il est donc possible d'étudier de manière fine les comportements des participants du marché.

3.3.3. Prix d'achat moyen et prix de revente moyen

Pour l'ensemble des propriétaires détenant un bien immobilier entre t et t + 1, les indicateurs $H_p(t)$ et $H_f(t)$ de la figure 7 rendent compte des prix moyens d'achat et de revente. Si on ne peut pas les assimiler entièrement à ces derniers, car il s'agit ici de moyennes géométriques pondérées par des mesures d'information, ils fournissent cependant des éléments d'analyse intéressants. Comme précédemment, ces grandeurs sont conditionnées par les bornes temporelles de l'échantillon : l'achat a lieu à partir de zéro et la revente au plus tard à T. Les graphiques ci-dessous présentent les évolutions de $H_p(t)$ et $H_f(t)$ en parallèle avec les indicateurs N-achat$_t$ et N-revente$_t$.

Figure 7 : Indicateurs de prix d'achat et de prix de revente

$H_p(t)$ et N-achat$_t$ $H_f(t)$ et N-revente$_t$



Courbes jaunes : incitations à l'achat et à la revente (échelles de droite)

$H_p(t)$ est croissant sur la période étudiée. Si dans la deuxième partie de l'intervalle [0,T] les prix de l'immobilier baissent fortement (figure 1), le phénomène

n'affecte cependant pas les prix d'achat moyens. L'indicateur N-achat$_t$ va permettre d'expliquer en partie ce phénomène. Le maximum des prix est atteint à la date t_{23} et dans la période suivante ceux-ci s'effondrent en passant de 190 à 166. Or, peu avant ce pic, on voit l'indicateur N-achat$_t$ baisser brutalement : à t_{22} il vaut 0,92 et oscille autour de 0,5 entre t_{23} et t_{26}. Les acheteurs ont donc significativement réduit leur activité avant le sommet de la courbe de prix, anticipant à juste titre que ceux-ci n'étaient pas soutenables. On retrouve ainsi, de manière observable, les éléments économiques fondamentaux inobservables, qui ont présidés au processus de génération des données. Au lieu de voir la courbe de $H_p(t)$ baisser, elle ne fait donc que rester constante car la réduction de l'activité d'achat a sous-pondéré les effets de cette chute.
Un autre élément permet d'expliquer ce phénomène de constance. Plus la date t est grande, plus la moyenne $H_p(t)$ porte sur un nombre important de dates d'achat. Les événements se produisant vers la fin de l'intervalle sont donc dilués dans les prix passés et par conséquent ils deviennent moins visibles dans la courbe $H_p(t)$.

L'allure de la courbe $H_f(t)$ est assez voisine de celle de l'indice, on retrouve en particulier le sommet des prix suivi de la chute. Même si l'incitation à la revente devient faible dans cette période, les prix de revente baissent car la moyenne $H_f(t)$ porte par définition sur les propriétaires qui auront revendu avant T. Ils ne pourront donc qu'être soumis à la baisse des prix et il n'y aura pas d'effets de dilution dans les valeurs passées.

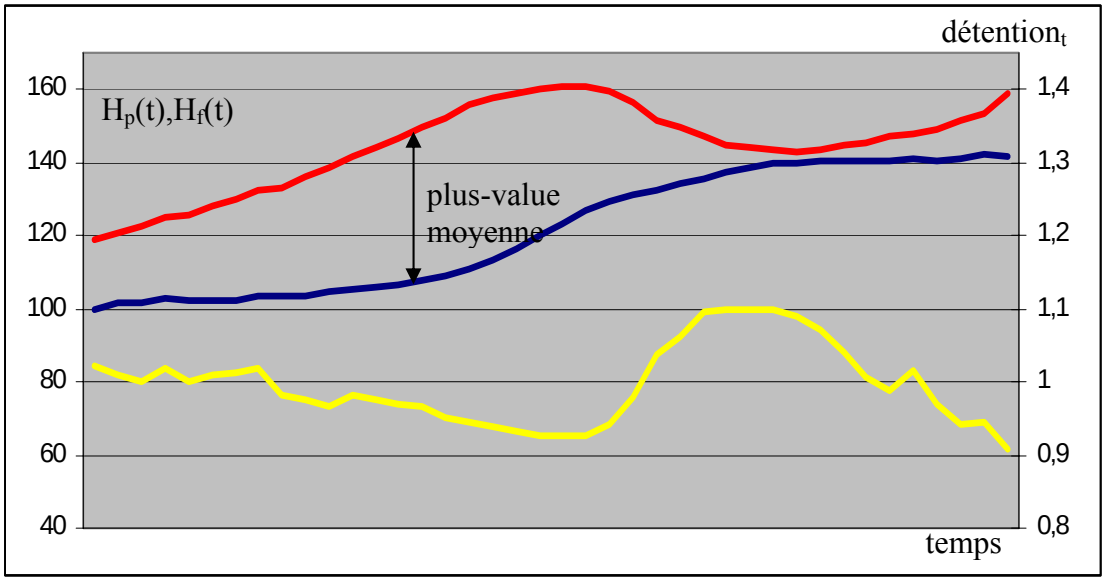
3.3.4. Durées de détention et rendements moyens

A l'image de N-revente$_t$ et N-achat$_t$, on peut construire un indicateur mesurant les variations de la période de détention en rapportant l'échantillon ω_0 à l'échantillon benchmark étalonné :

$$\text{détention}_t = \tau^t(\text{échantillon } \omega_0) / \tau^t(\text{échantillon étalonné}) \qquad (15)$$

Une valeur supérieure à 1 détecte un rallongement de la période de détention, une valeur inférieure à 1 un raccourcissement. Le graphique ci-dessous met en parallèle les prix moyens d'achat et de revente, $H_p(t)$ et $H_f(t)$, et ce nouvel indicateur. Toutes ces grandeurs sont observables. Dans la première partie $détention_t$ baisse progressivement jusqu'à la date t_{21}, il commence à remonter dès t_{22} et t_{23} (sommet des prix, cf. figure 1) pour atteindre son niveau maximal pendant la chute des prix. Vers la fin de l'intervalle cet indicateur revient autour de la valeur pivot 1 (la dernière valeur de 0,9 est sans doute la conséquence d'un effet de bord).

Figure 8 : prix moyens d'achat et de revente, indicateur de durée de détention



Courbe bleue : $H_p(t)$ Courbe rouge : $H_f(t)$ Courbe jaune : $détention_t$

L'écart entre la courbe bleue (prix moyen d'achat) et la courbe rouge (prix moyen de revente) s'interprète comme la plus-value moyenne réalisée par les propriétaires qui possédaient de l'immobilier à la date considérée. Au début de l'intervalle, ce gain est approximativement constant et la durée de détention est à un niveau standard. Progressivement, l'écart entre les deux courbes s'accroît et parallèlement la durée de détention se raccourcit : la forte augmentation des prix dans cette période a incité les propriétaires à revendre plus rapidement pour profiter de cette situation favorable. Dans un second temps, la situation s'inverse nettement.

L'écart se resserre, indiquant une réduction des plus-values, et la durée de détention s'allonge significativement, témoignant de la volonté des vendeurs de repousser la date de revente en espérant une situation plus propice[28].
Un cas de figure intéressant aurait pu également se présenter : celui du croisement des deux courbes $H_f(t)$ et $H_p(t)$. Si pendant quelques dates $H_p(t)$ est au-dessus de $H_f(t)$, les propriétaires possédant de l'immobilier à ces instants réaliseront, en moyenne, une moins-value sur leur investissement. Une telle situation sur le marché de l'immobilier résidentiel ne serait certainement pas sans conséquences[29] pour son évolution future.

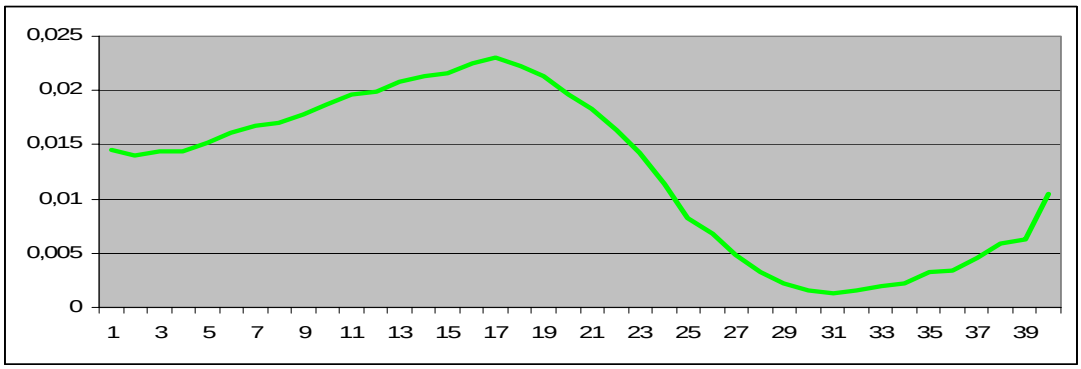
Les trois courbes de la figure 8 permettent ainsi d'étudier directement les aléas économiques et les réactions comportementales des propriétaires. Il faut remarquer ici qu'il n'a pas été nécessaire de collecter de nouvelles données ou d'élaborer de nouvelles méthodologies économétriques pour cela. L'emploi des différents concepts introduits lors de la phase de reformulation et de décomposition théorique de l'indice de ventes répétées a suffit. L'extraction de l'information contenue dans l'échantillon d'estimation s'est accrue donc très sensiblement, les résultats du RSI ne se résumant plus maintenant aux seules valeurs indicielles.

Le taux moyen $\rho_t = (1 / \tau^t) * (\ln H_f(t) - \ln H_p(t))$ est également un outil d'analyse intéressant, ses variations sont illustrées avec la figure 9. Au début de l'intervalle, sous le double effet de l'augmentation de la plus-value et de la baisse de la durée de détention, il s'élève significativement. Puis dans un deuxième temps, $H_p(t)$ et $H_f(t)$ devenant très proches, il devient presque nul. Pour les propriétaires engagés à ces dates, la rentabilité de leur investissement est donc en moyenne nulle. Une telle situation laisse probablement de mauvais souvenirs à un particulier... Dans

[28] Ces deux phénomènes sont directement dus à l'influence de la variable INC dans le processus de génération des données. Cette grandeur fondamentale est inaccessible à l'observateur en pratique ; par contre l'étude des indicateurs observables que sont $H_p(t)$, $H_f(t)$ et $detention_t$ permet d'en retrouver les effets.
[29] On pourrait par exemple imaginer une migration du statut de propriétaire au statut de locataire pour une partie significative des agents ayant vécu cet épisode.

l'éventualité où la rentabilité deviendrait négative, il serait même raisonnable de penser qu'elle pourrait alors engendrer un a priori défavorable pour les futurs achats immobiliers, créant ainsi un assèchement relatif du marché (du côté de la demande).

Figure 9 : Moyenne des taux moyens ρ_t



4. L'étude des sensibilités de l'indice

On définit dans ce paragraphe divers indicateurs destinés à mesurer la fiabilité de l'indice. Puis, à partir du procédé de génération d'échantillons présenté dans le paragraphe 2, en modifiant adéquatement certaines conditions de simulation, on étudiera la sensibilité du RSI aux paramètres du modèle.

4.1. Mesures de performance pour un indice

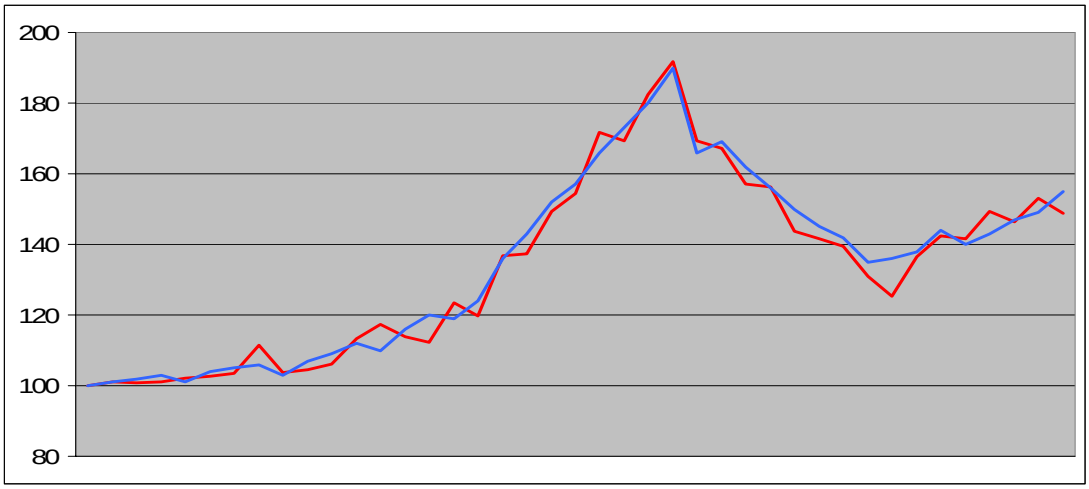
Le but d'un indice immobilier est d'approcher à chaque date le « vrai prix de l'immobilier », que l'on a défini comme la moyenne des prix de tous les biens existants[30]. Cette courbe de référence est en pratique inconnue et la notion de "vrai

[30] Et non pas des seuls biens échangés à la date t car on ne construit alors qu'une mesure sur le flux et pas une mesure sur le stock, cf. note 2 de ce chapitre.

prix" peut-être même inexistante, mais pour tester la fiabilité du modèle on la suppose donnée et fixée.

Les indices sont estimés pour les dates de 1 à T, en prenant une base 100 à t = 0. La figure ci-dessous illustre le type de résultats que l'on peut obtenir. La courbe bleue correspond au « vrai prix » et la rouge à celle obtenue par la méthode des ventes répétées[31], à partir d'un échantillon simulé ω_0. Le résultat peut sembler convenable graphiquement mais on ne peut cependant pas se contenter d'estimer l'adéquation des deux courbes visuellement, il faut pouvoir construire des indicateurs de performance plus précis.

Figure 10 : Vrai prix de l'immobilier et résultat de l'indice de ventes répétées



Si l'on souhaitait réaliser une modélisation économique classique, les prix du foncier devraient être décrits par un processus aléatoire X(ώ,t) pour rendre compte du risque fondamental de l'économie. La courbe bleue serait dans ce cas une trajectoire, c'est-à-dire une réalisation du processus pour un état de la nature $\acute{\omega}_0$ représentant le fonctionnement de l'économie dans son ensemble. Comme le but poursuivi ici n'est pas de travailler sur le risque immobilier lui-même mais simplement sur la fiabilité de l'indice, on est amené à modifier ce formalisme. On supposera que l'état de la

[31] La démarche permettant d'obtenir ce genre de graphique a été présentée en détail dans le paragraphe précédent.

nature $\acute{\omega}_0$ est connu, ce qui revient en fait à considérer que la courbe bleue dépend uniquement du temps. Le phénomène aléatoire concernera l'échantillon d'estimation observé, sachant le contexte économique global. La courbe rouge, quant à elle, dépendra du temps et d'un alea traduisant le phénomène de fluctuation de l'échantillon.

Pour la date t, on notera par J(t) la fonction déterministe du « vrai prix » et par RSI(t,ω) la valeur fournie par l'indice estimé, pour une situation économétrique correspondant à un lot de données ω. L'erreur commise en pourcentage, 100*(RSI(t, ω) / J(t) – 1), sera notée e(t, ω). Plusieurs indicateurs mesurant l'écart entre les deux courbes peuvent alors être définis

- $\text{BIAIS}(\omega) = \left(\sum_{t=1,\ldots,T} e(t, \omega) \right) / T$

  Pour un jeu de données particulier ω, cette grandeur représente l'erreur moyenne algébrique entre la courbe de référence et l'indice obtenu avec cet échantillon. Le signe de l'erreur étant pris en compte il est alors possible de savoir si, pour ce jeu de données particulier, l'indice a surévalué ou sous-évalué les prix de l'immobilier.

- $\text{ERREUR\_MOYENNE}(\omega) = \left(\sum_{t=1,\ldots,T} |e(t, \omega)| \right) / T$

  Pour un échantillon ω, cette grandeur représente l'erreur moyenne absolue. Elle ne peut pas servir à la détection d'éventuels biais, sa fonction est simplement de mesurer la qualité de l'ajustement.

- $\text{ECART-TYPE}(\omega) = \left[\sum_{t=1,\ldots,T} (|e(t, \omega)| - \text{ERREUR\_MOYENNE}(\omega))^2 / T \right]^{1/2}$

  Écart-type de l'erreur absolue pour un échantillon particulier ω. Cette grandeur mesure la dispersion autour de ERREUR_MOYENNE(ω) pour la distribution des erreurs absolues $|e(t, \omega)|$.

- ERREUR_MAX(ω) = $\underset{t=1,\ldots,T}{\text{Max}} |e(t, \omega)|$

  Distance maximale observée dans la simulation ω entre les deux courbes

Le tableau suivant présente un exemple des résultats que l'on peut obtenir, pour deux échantillons Ω_1 et Ω_2, générés grâce au processus décrit précédemment dans le paragraphe 2, sous des conditions de simulations[32] vraisemblables.

Tableau 4 : Deux exemples de simulation

	BIAIS(ω)	ERREUR_MOYENNE(ω)	ECART-TYPE(ω)	ERREUR_MAX(ω)
Ω_1	3.01	3.01	2.29	9.29
Ω_2	- 1.63	1.63	1.36	5.39

Comme on peut le constater les résultats varient en fonction du jeu de données, en raison du phénomène de fluctuation d'échantillonnage. Afin d'avoir une perception globale et stable de la qualité de l'indice il est alors nécessaire de travailler sur les espérances empiriques de ces variables. Pour cela on simulera un certain nombre de lots de données (en général une centaine permettra d'atteindre un niveau de régularité suffisant) et on calculera des mesures globales de performance :

- BIAIS = E [BIAIS(ω)]

  Moyenne des biais obtenus sur un grand nombre d'échantillons. Il est beaucoup plus significatif que BIAIS(ω), calculé sur un seul jeu. Si sa valeur est sensiblement différente de zéro, il témoigne d'un problème méthodologique.

- ERREUR_MOYENNE = E [ERREUR_MOYENNE(ω)]

  Moyenne des erreurs moyennes sur un grand nombre d'échantillons

[32] Les conditions de simulation sont identiques pour Ω_1 et Ω_2

- ECART-TYPE_MOYEN = $E\,[\text{ECART-TYPE}(\omega)]$

  Moyenne des écart-types sur un grand nombre d'échantillons

- ERREUR_MAX_MOYENNE = $E\,[\text{ERREUR\_MAX}(\omega)]$

  Moyenne des erreurs maximales observées sur un grand nombre d'échantillons

- MAX_ ERREUR_MAX = $\text{Max}_{\omega}\,[\text{ERREUR\_MAX}(\omega)]$

  Plus grande erreur observée sur l'ensemble des échantillons

Sous les mêmes conditions de simulation que dans le tableau précédent, les résultats obtenus pour les indicateurs agrégés sont présentés ci-dessous, en fonction du nombre de scénarios générés.

Tableau 5 : Indicateurs agrégés en fonction du nombre de simulations

Nombre de simulations	BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	ERREUR_MAX MOYENNE	MAX ERREUR_MAX
10	0.39	2.88	1.88	7.54	9.44
50	0.18	2.84	1.95	7.60	14.14
100	0.15	2.90	1.92	7.88	14.95
1000	0.13	2.91	1.88	7.73	17.14

Avec 100 simulations on atteint une certaine stabilité pour les quatre premiers indicateurs, on considérera donc qu'il s'agit du niveau standard requis. Si la variabilité des échantillons augmente, on pourra augmenter en parallèle le nombre des simulations pour retrouver un niveau de stabilité suffisant. Mais, aller jusqu'à 1000 de façon systématique poserait des problèmes de temps de calcul.

Le cinquième indicateur renseigne sur la plus grande valeur observée pour l'ensemble des jeux de données et l'ensemble des dates. Il s'agit d'un indicateur de valeur extrême. Comme par définition les événements extrêmes sont rares, il est normal d'avoir une certaine instabilité de cette mesure. Afin d'affiner la connaissance des queues de distribution des erreurs, on pourra également introduire des concepts de type VaR si le besoin s'en fait sentir. Pour cela on génèrera un lot d'échantillons suffisamment grand, les erreurs observées seront ensuite rangées par ordre croissant et on pourra alors déterminer différents quantiles :

- $Q_{10\%}$: 90% de la distribution des erreurs est sous ce seuil
- $Q_{5\%}$: 95% de la distribution des erreurs est sous ce seuil
- $Q_{1\%}$: 99% de la distribution des erreurs est sous ce seuil

Pour 100 scénarios, sous les mêmes conditions que ci-dessus, on obtient ainsi :

$Q_{10\%} = 6.11$ $Q_{5\%} = 7.09$ $Q_{1\%} = 9.38.$

Enfin, il faut remarquer que tous ces indicateurs agrègent des erreurs de dates différentes. Dans certains cas on sera amené à les décomposer en sous-indicateurs pour étudier la dimension temporelle de l'erreur.

4.2. La sensibilité de l'indice au contexte économique

Il est a priori possible que la technique des ventes répétées se comporte mieux dans certains contextes économiques que dans d'autres. La tendance globale (marché stable, haussier ou baissier) et la volatilité des prix de l'immobilier pourraient par exemple avoir un impact sur la qualité de l'indice. En termes de modélisation, cette question revient à examiner l'influence éventuelle des moments d'ordre 1 et 2 des taux de croissance monopériodiques de la courbe des « vrais prix » sur les mesures d'erreurs.

Pour étudier ce problème, on générera différents échantillons associés à des courbes de référence[33] simples et variées, tout en gardant les autres paramètres du modèle inchangés. On reprend pour cela le processus de génération des échantillons décrit pour le scénario de référence (paragraphe 2). On supposera cette fois que les niveaux d'activité du marché sont toujours constants ($K_i = 1$ et $\mathcal{K}_0 = 10000$). Le temps d'égalité des bruits Θ sera fixé à 10. Les dates de reventes seront simulées à partir des taux de survie[34] mais en supprimant l'effet de la variable INC[35] pour ne conserver que l'aléa. Les valeurs des moyennes à l'achat et à la vente dans chaque classe de ventes répétées, $h_p^{(i,j)}$ et $h_p^{(i,j)}$, seront déduites des niveaux des courbes de référence en perturbant ceux-ci par un coefficient multiplicatif suivant une loi log-normale de paramètres $\mu = 0$ et $\sigma = 0.05$, afin de simuler le phénomène de fluctuation d'échantillonnage pour les prix[36].

4.2.1. L'impact de la tendance

Afin d'évaluer la sensibilité de l'indice à la tendance, on suppose que la courbe de référence est exponentielle, d'équation $J(t) = 100\exp(\lambda t)$. Le paramètre λ s'interprète comme le taux de variation des prix immobiliers d'une période sur l'autre, il est supposé constant (si λ est nul le marché est stable, haussier si $\lambda > 0$ et baissier sinon). Pour chaque valeur de λ on simulera 100 scénarios et on calculera les valeurs de différents indicateurs d'erreurs, les résultats sont présentés dans le tableau 6.

L'indicateur de biais ne dépasse pas deux millièmes et son signe n'est pas constant. On peut donc affirmer qu'il n'y a pas de surestimation ou de sous-

[33] Courbe des « vrais prix ».
[34] Taux de survie entre t et t+1 = 1 – (0,03 + INC + ε).
[35] L'effet de INC sera étudié plus en détail par la suite.
[36] On a par exemple $h_p^{(i,j)}$ = (niveau à t_i de la courbe de référence) x ε , où ε suit la loi indiquée ($E[\varepsilon] \approx 1$).

estimation systématique causée par la tendance. Comme de plus les quatre autres indicateurs ne semblent pas non plus varier significativement en fonction du contexte économique, on peut considérer que la tendance globale du marché n'affecte pas la qualité de l'indice.

Tableau 6 : Effet de la tendance sur l'indice

λ	BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	ERREUR_MAX MOYENNE	MAX ERREUR_MAX
-0.15	-0.13	1.42	0.94	3.86	6.46
-0.1	0.01	1.62	1.01	4.18	7.70
-0.05	0.01	1.48	0.97	3.99	7.25
0	-0.09	1.56	0.97	4.05	7.68
0.05	0.12	1.53	0.96	3.93	6.25
0.1	0.16	1.58	0.99	4.11	6.79
0.15	0.04	1.59	0.97	4.14	8.84

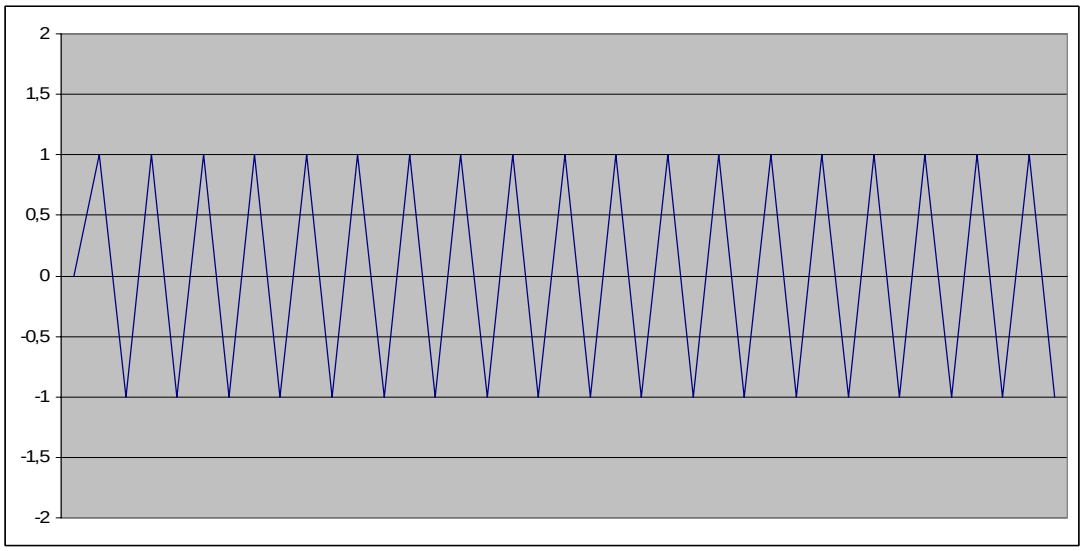
4.2.2. L'impact de la volatilité des prix immobiliers

Lorsqu'un marché traverse une période d'incertitude forte, cela peut se traduire par une volatilité importante. Si l'indice de ventes répétées induit indirectement un lissage des prix, il pourrait alors avoir du mal à reproduire ces phénomènes d'oscillation et il sous-estimerait par conséquent le risque de l'investissement immobilier. Pour tester sa réaction, on construit un signal oscillant de base, noté f(t), prenant alternativement les valeurs +1 et -1 et dont l'écart-type vaut 1 (figure 11).

A partir d'une situation sans volatilité, où la fonction J du « vrai prix » est constante et vaut 100 à chaque date, on introduit progressivement de la volatilité en

multipliant J(t) par (1 + α f(t) / 100). Si le paramètre α est par exemple fixé à 5, la courbe que l'on obtient prend alternativement les valeurs 105 et 95. Ce nombre α correspond simplement à l'écart-type du nouveau signal. Comme précédemment, on génère 100 scénarios pour différentes valeurs de α. Les niveaux atteints par les indicateurs d'erreurs sont présentés dans le tableau 7.

Figure 11 : Signal de base



Ces résultats indiquent une absence de biais et une grande stabilité des indicateurs ERREUR MOYENNE et ECART-TYPE MOYEN. Les deux dernières colonnes donnent aussi des résultats stables, mais un peu moins que ceux des colonnes deux et trois. Cette particularité est liée à la nature de ERREUR_MAX MOYENNE et MAX ERREUR_MAX qui sont en effet des mesures de valeurs extrêmes. Pour stabiliser ces indicateurs il faudrait simuler un plus grand nombre de scénarios. Toutefois, comme la variabilité obtenue reste cependant raisonnable, l'analyse ne sera pas approfondie dans cette direction. Globalement la volatilité de la courbe de référence n'a donc pas d'incidence sur la fiabilité de l'indice.

Tableau 7 : Effet de la volatilité des prix immobiliers sur la qualité de l'indice

α	BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	ERREUR_MAX MOYENNE	MAX ERREUR_MAX
0	-0.10	1.43	0.94	3.84	7.07
5	-0.05	1.52	0.96	4.03	6.52
10	-0.02	1.54	0.96	3.96	7.48
20	-0.19	1.49	0.96	3.97	6.69
30	0.04	1.50	0.96	3.99	6.71

4.2.3. Conclusion

Les simulations développées ci-dessus correspondent à des situations très marquées. $\lambda = 0.05$ signifie par exemple que les prix partent d'un niveau 100 à t_0 pour aboutir à un niveau de 739 à t_{40} et une valeur de 20 pour α produit des prix immobiliers oscillant entre 120 et 80, de période en période. Or, malgré ces valeurs caricaturales, l'indice de ventes répétées reste très stable. Les phénomènes observés dans la réalité étant en général de plus faible ampleur que ceux simulés ici, on peut donc raisonnablement en conclure que la qualité de l'indice n'est pas affectée par le contexte économique, que ce soit en tendance ou en volatilité.

4.3. La sensibilité de l'indice au temps d'égalité des bruits

Dans le modèle de Case-Shiller, la variance du résidu $2\sigma_N^2 + \sigma_G^2 (t_j(k) - t_i(k))$ se compose de deux termes. Le premier, $2\sigma_N^2$, capture les imperfections inhérentes au marché et supposées constantes dans le temps (modélisation par un bruit blanc), tandis que le $\sigma_G^2 (t_j(k) - t_i(k))$ cherche à capturer la composante temporelle du bruit (modélisation par une marche aléatoire gaussienne). Une vente répétée consistant en

un achat et une revente les imperfections du premier type peuvent s'exprimer deux fois, d'où le 2 devant σ_N^2. La deuxième partie de la variance est, quant à elle, directement proportionnelle au temps de détention $(t_j(k) - t_i(k))$; plus celui-ci est long, moins la vente répétée sera informationnelle pour l'indice. Dans la partie théorique, le paramètre $\Theta = 2\sigma_N^2/ \sigma_G^2$ est apparu dans les équations lors de la généralisation du modèle. Il mesure l'importance relative des perturbations et peut s'interpréter comme le temps d'égalité des deux sources de bruits. Pour des durées de détention inférieures à Θ, le bruit blanc domine dans la variance du résidu mais pour des durées supérieures à Θ la marche aléatoire constitue la source principale du bruit. Le modèle a d'abord été développé dans le paragraphe 3 du chapitre 1 en prenant $\Theta = 0$ (absence d'imperfections, tendances locales conservées). Les articles empiriques fournissent des valeurs entre 4 et 25 pour ce paramètre, et si l'on choisit un Θ voisin de $+\infty$ on retrouve le modèle BMN. En effet, $\Theta \rightarrow +\infty$ correspond à un σ_N^2 très supérieur à σ_G^2 ; en d'autres termes la marche aléatoire est négligeable dans la variance du résidu, seule subsiste la partie constante $2\sigma_N^2$ comme dans Bailey, Muth et Nourse (1963).

Afin d'étudier le comportement des indicateurs d'erreurs en fonction de Θ, on simule pour chaque valeur de ce paramètre une centaine de jeux de données. Le niveau de la courbe de référence est supposé constant[37] (100 pour chaque date) ainsi que l'activité du marché (K_i =1 et $\mathcal{K}_0$ = 10000). La variable INC est neutralisée et les moyennes géométriques $h_p^{(i,j)}$ et $h_f^{(i,j)}$ sont obtenues en perturbant la courbe de référence par une log-normale de paramètres $\mu = 0$ et $\sigma = 0{,}05$. Les résultats sont présentés dans le Tableau 8.

Les indicateurs ERREUR_MOYENNE et ECART-TYPE_MOYEN varient à l'opposé de Θ et semblent atteindre pour $\Theta \rightarrow +\infty$ des valeurs seuils

[37] On a pu vérifier dans le paragraphe précédent que le contexte économique (en tendance et en volatilité) n'avait pas d'impact significatif sur la fiabilité de l'indice. D'une manière générale, on prendra donc dorénavant dans les études de sensibilité, une situation économiquement neutre comme référence.

approximativement égales à la moitié de celles obtenues pour $\Theta = 0$. Le biais détecté reste faible, on peut cependant constater qu'il est légèrement supérieur (en valeur absolue) pour les premières valeurs du paramètre. Si l'on compare un modèle Case-Shiller classique ($\Theta = 10$) et un modèle BMN ($\Theta \rightarrow +\infty$), les indicateurs d'erreurs du premier cas sont supérieurs d'approximativement 20% à ceux du modèle BMN. Néanmoins, ce résultat ne signifie pas que le choix d'une spécification hétéroscédastique (CS) pour la variance du résidu est mauvais, car il faut aussi tenir compte du processus de génération des échantillons. En effet les moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$, pour une classe (i,j) de ventes répétées, sont simulées grâce à des perturbations aléatoires, à variances constantes, des vrais prix. Elles sont donc indépendantes de la longueur de la période de détention j – i, à l'image des prix dans BMN. Le modèle BMN est donc plus adapté par nature à l'échantillon généré. En revanche, comme le modèle Case-Shiller est destiné à des données présentant une variance croissante avec le temps de détention (phénomène absent des échantillons de simulation), il n'est pas surprenant de détecter des erreurs plus grandes pour les premières valeurs de Θ.

Tableau 8 : Indicateurs d'erreurs en fonction du temps d'égalité des bruits (échantillon de type BMN)

Θ	BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	ERREUR_MAX MOYENNE	MAX ERREUR_MAX
0	0.38	2.33	1.49	6.01	10.97
2	-0.20	1.79	1.13	4.67	8.39
5	-0.20	1.67	1.04	4.37	7.89
10	0.30	1.63	1.01	4.12	8.00
15	-0.01	1.40	0.93	3.74	6.04
20	-0.02	1.47	0.93	3.80	6.97
30	-0.14	1.43	0.93	3.77	7.17
50	-0.08	1.36	0.88	3.63	5.88
100	0.13	1.32	0.87	3.54	5.95
1000	0.00	1.37	0.86	3.59	5.58
10000	0.04	1.35	0.86	3.56	6.37
100000	-0.17	1.36	0.86	3.50	6.86
1000000	0.16	1.35	0.84	3.53	7.54

Afin de tester la réaction des erreurs aux valeurs de Θ pour des échantillons de type Case-Shiller les simulations sont réitérées en conservant les mêmes paramètres, à l'exception des écarts-types pour les moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$. La volatilité dépendra ici de la classe de ventes répétées, et donc de la période de détention[38] via la formule $\sigma^{(i,j)} = 0{,}025 + 0{,}04((j - i)/40)^{½}$. Les moyennes pourront donc ici davantage s'écarter des vrais prix pour les détentions longues, à l'image de ce qui se passe pour les $p_{k,i}$ dans le modèle CS. Les résultats sont présentés dans le tableau 9.

Il n'est pas vraiment possible de comparer les niveaux absolus des indicateurs entre ces deux types d'échantillons. On peut cependant constater que les valeurs asymptotiques[39] pour les deux principaux indicateurs, ERREUR_MOYENNE et ECART-TYPE_MOYEN, sont très voisines. Dans le premier cas la stabilisation des erreurs intervient vers Θ = 40 tandis qu'elle se produit dès Θ = 10 dans le second.

Tableau 9 : Indicateurs d'erreurs en fonction du temps d'égalité des bruits (échantillon de type Case-Shiller)

Θ	BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	ERREUR_MAX MOYENNE	MAX ERREUR_MAX
0	-0.20	1.81	1.12	4.64	9.22
2	-0.40	1.60	0.98	3.99	7.21
5	-0.05	1.41	0.87	3.59	8.88
10	0.09	1.37	0.86	3.55	6.77
15	-0.09	1.39	0.85	3.53	5.95
20	0.08	1.38	0.84	3.48	5.76
30	0.00	1.37	0.86	3.51	6.95
50	-0.21	1.34	0.84	3.50	6.76
100	0.05	1.33	0.83	3.44	7.32
1000	0.13	1.38	0.85	3.49	7.17
10000	0.12	1.35	0.86	3.54	5.93
100000	0.13	1.34	0.84	3.44	5.85
1000000	0.12	1.33	0.83	3.43	6.64

[38] Pour les détentions courtes (j – i = 1) la volatilité sera de 0,031 et pour les détentions longues (j – i = 40) de 0,065

[39] $\Theta \rightarrow +\infty$

Si pour un échantillon de type BMN il est cohérent de trouver des valeurs d'erreurs minimales quand le modèle est à variance constante ($\Theta \rightarrow +\infty$), on aurait pu penser que dans le cas des données à variance hétéroscédastique les indicateurs d'erreurs auraient atteint un minimum pour une valeur finie de Θ, correspondant au Θ réel associé à l'échantillon. Or ce n'est pas ce que l'on constate. On peut toutefois conjecturer que pour un Θ égal à cette valeur, la stabilisation des erreurs est réalisée. D'une manière générale, pour pouvoir analyser plus en détail l'impact de Θ, il faudrait utiliser des échantillons où l'on simulerait directement les prix $p_{k,i}$ et $p_{k,j}$, et non plus seulement les grandeurs intermédiaires $h_p^{(i,j)}$ et $h_f^{(i,j)}$.

En résumé, le temps d'égalité des bruits à un impact sur les erreurs. Pour les indicateurs étudiés ici le modèle de Case-Shiller ne démontre pas, sur les données simplifiées de la simulation, une efficacité supérieure au modèle BMN. Cependant on ne peut pas en conclure qu'il lui est inférieur, car les erreurs se stabilisent rapidement dans le cas des données à variance non constante.

4.4. La sensibilité de l'indice à la volatilité des prix

On simule pour différents niveaux de σ (paramètre de volatilité commun aux moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$) 500 scénarios pour lesquels $\Theta = 10$, la liquidité est constante ($K_i = 1$ et $\mathcal{K}_0 = 10000$) et la courbe des vrais prix est plate. Les valeurs obtenues pour les différents indicateurs sont présentées dans le tableau 10.

La méthode de génération des données développée dans ce travail ne permet pas d'étudier directement l'impact de la volatilité des prix sur la fiabilité de l'indice car on ne simule pas chaque transaction en détail, comme mentionné précédemment. L'étude partielle de l'influence de ce paramètre est cependant rendue possible par l'intermédiaire de $h_p^{(i,j)}$ et $h_f^{(i,j)}$. La volatilité d'une moyenne étant plus faible que la

volatilité de ses constituants[40] on couvre facilement l'ensemble des volatilités raisonnables pour les prix en utilisant le paramètre σ. Ainsi, le niveau maximal de 0,5 pour σ produit des fluctuations de la moyenne $h_p^{(i,j)}$ autour de sa valeur centrale pouvant aller jusqu'à +/- 50%. Un tel niveau de dispersion étant déjà très suspect pour les prix, il l'est à fortiori encore plus pour une moyenne de prix, car il supposerait implicitement une variabilité démesurée sur les transactions réelles. Le tableau 10 permet donc de mesurer indirectement l'ampleur des perturbations résultant de la variance réelle des prix.

Tableau 10 : Indicateurs d'erreurs en fonction de la volatilité des moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$

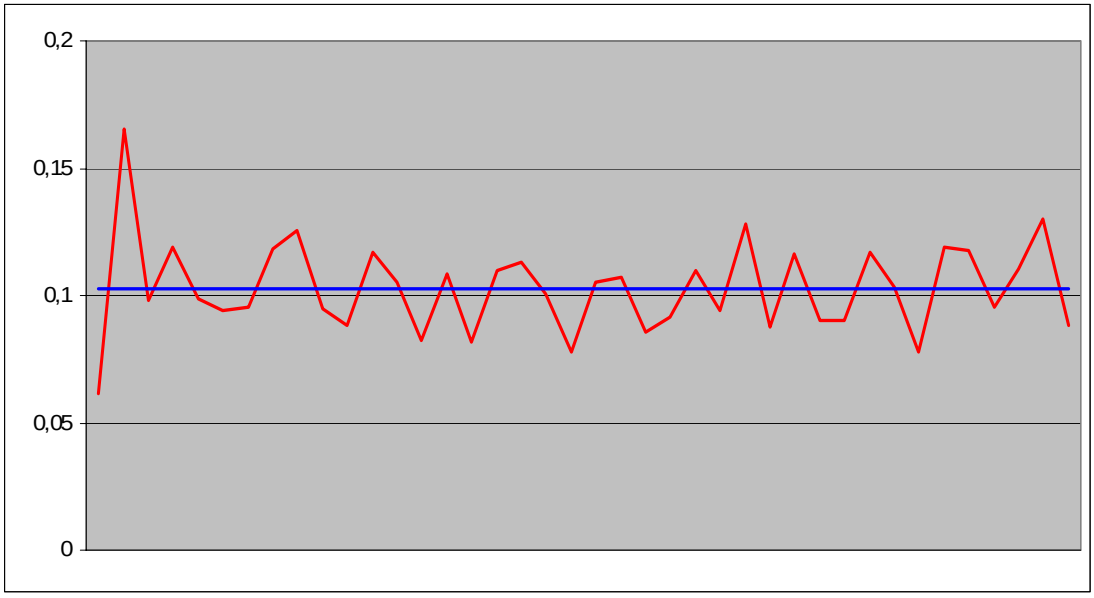
σ	BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	ERREUR_MAX MOYENNE	MAX ERREUR_MAX
0.01	-0.02	0.31	0.20	0.80	1.64
0.02	-0.01	0.62	0.39	1.61	3.18
0.05	0.04	1.53	0.97	3.98	9.94
0.10	0.10	3.67	2.35	9.67	23.12
0.15	0.22	4.58	2.94	12.10	27.51
0.2	0.50	6.20	3.95	16.22	43.79
0.3	0.77	9.39	5.99	24.71	64.64
0.4	1.00	12.30	7.95	32.84	80.51
0.5	1.84	15.70	10.12	42.10	122.98

A l'analyse des résultats on constate, comme on pouvait s'y attendre, que plus la volatilité augmente moins l'indice est précis. Toutefois même pour des niveaux très élevés de volatilité, l'erreur moyenne reste raisonnable (inférieure à 10% pour σ = 0.3 par exemple), en raison de la taille importante des échantillons d'estimation.

[40] s'ils sont indépendants et identiquement distribués

L'autre élément notable est l'apparition d'un léger biais positif. On redémontrera dans le chapitre 3, paragraphes 2.6.4 et 2.6.5, que le RSI est très faiblement biaisé[41] et une formule théorique quantifiant précisément ce décalage sera établie. A l'image de l'indicateur ERREUR_MOYENNE, dont on constatera dans le paragraphe 5.2.1 qu'il présente une forme en U quand on le décompose sur [0,40], on peut se demander si le biais présente une structure temporelle particulière. Pour cela, on désagrège la mesure globale BIAIS sur [0,40] dans la simulation. En choisissant $\sigma = 0{,}10$ et en générant 25000 scénarios les résultats, illustrés dans la figure 12, semblent être homogènes (à l'exception éventuellement de la deuxième date). Le biais est donc à priori uniformément distribué sur [0,40]. On verra toutefois, lors de l'étude théorique, que ce propos devra être nuancé.

Figure 12 : Evolution du biais sur [0,40] pour $\sigma = 0{,}10$ (25000 scénarios)



[41] résultat déjà mentionné par Goetzmann (1992).

5. Fiabilité de l'indice et distribution des ventes répétées

Une base de données étant rarement homogène, il peut arriver qu'il y ait des périodes avec très peu de ventes répétées, voire des trous si la saisie des informations dans la base a été interrompue pendant quelques temps. Corrélativement, on peut aussi être confronté à des échantillons où l'information est concentrée sur quelques sous-périodes. Avant d'étudier les impacts de l'hétérogénéité de la distribution des $\{n_{i,j}\}$ sur l'indice, on étudiera préalablement un autre problème, à savoir celui de la fiabilité de l'indice pour les échantillons de petite taille. La troisième section sera consacrée à l'étude d'une caractéristique importante de la distribution des ventes répétées, et qui expliquera certains des phénomènes curieux que l'on rencontrera dans ce paragraphe : l'asymétrie des données.

5.1. L'impact de la taille de l'échantillon

5.1.1. Qu'est-ce qu'un petit échantillon ?

La courbe des « vrais prix » est supposée constante, égale à 100 pour toutes les dates de l'intervalle [0,40]. Le temps d'égalité des bruits Θ est fixé à 10 et les moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$ sont déduites de la courbe de référence en perturbant celle-ci, pour chaque date, par un coefficient multiplicatif suivant une loi log-normale de paramètres $\mu = 0$ et $\sigma = 0{,}05$. Dans le scénario présenté dans le paragraphe 3 de ce chapitre le taux de survie se calculait de manière générale par :

$$\text{Taux de survie entre t et t+1} = 1 - (0{,}03 + \text{INC} + \varepsilon)$$

Très classiquement, nous supposerons ici que la variable INC est nulle, afin de ne pas introduire d'interférences avec le phénomène étudié. On adoptera de plus la

simulation binomiale pour pouvoir générer des échantillons de très petite taille (cf. paragraphe 2.4).

La série des $\{K_i\}$, nombre de biens échangés sur le marché à chaque date, est la variable de référence pour étudier la sensibilité de l'indice à la taille de l'échantillon (noté N dans la partie théorique). Si l'on suppose la série des $\{K_i\}$ constante, elle est alors entièrement déterminée par la valeur $\mathcal{K}_0$. L'intervalle temporel étudié ici, [0,40], génère (40 x 41) / 2 = 820 possibilités pour le couple (date d'achat ; date de revente). Un échantillon de 300 données signifiera donc que les 2/3 des $n_{i,j}$ sont nuls (en moyenne). Le tableau 11 présente les résultats obtenus en simulant une centaine de lots de données, pour les indicateurs d'erreurs BIAIS, ERREUR_MOYENNE et ECART-TYPE_MOYEN, en fonction des valeurs de $\mathcal{K}_0$; la dernière colonne donnant la taille moyenne N des échantillons générés.

Le biais apparaît négligeable et non systématique. Les valeurs asymptotiques pour les très grands échantillons sont approximativement de 1.55 et 1.00 pour ERREUR_MOYENNE et ECART-TYPE_MOYEN. A partir de $\mathcal{K}_0$ = 500, les résultats semblent osciller autour de ces valeurs limites. On pourra donc considérer que cette taille d'échantillon est suffisante pour rendre l'indice fiable. En des termes plus concrets, quand $\mathcal{K}_0$ = 500 le nombre de ventes répétées est en moyenne de 8600. Si on le rapporte au nombre de types de ventes répétées, 820, il faut donc en moyenne une dizaine d'éléments dans chaque classe pour atteindre la meilleure qualité possible. D'une manière générale on proposera donc l'heuristique suivante :

> *« Pour l'étude d'un intervalle [0,T], la taille minimale de l'échantillon[42] doit être de 5T(T + 1). Pour les valeurs inférieures la fiabilité de l'indice sera réduite en raison de la rareté des données, on parlera dans ce cas de petit échantillon »*

[42] 10 x [nombre de couples (date d'achat ;date de revente) possibles] = 10 [T (T+1) / 2] = 5T(T+1).

Ce seuil est valable pour des données assez homogènes, et sous les conditions employées dans la simulation, en particulier $\sigma = 0{,}05$. Le problème des échantillons comportant des zones à forte densité de données et d'autres avec des densités beaucoup plus faibles sera examiné dans la section suivante.

Tableau 11 : Indicateurs d'erreurs et taille moyenne de l'échantillon en fonction K_0

K_0	BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	N
5	-0.09	5.17	3.53	86
10	0.04	3.37	2.22	172
20	-0.15	2.68	1.67	342
50	0.15	1.93	1.28	861
100	0.00	1.77	1.13	1 726
500	-0.09	1.55	0.99	8 599
1 000	-0.05	1.63	1.02	17 235
10 000	-0.01	1.51	0.97	172 217
100 000	0.03	1.58	0.99	1 722 576

5.1.2. Les erreurs extrêmes pour les petits échantillons

En analysant plus en détails la méthodologie et les résultats présentés dans le tableau 11, on peut penser que les indicateurs sous-estiment probablement la taille des erreurs pour les petits échantillons. En effet, l'écart-type des moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$ a été fixé à $\sigma = 0{,}05$, indépendamment de la valeur de N. Or pour les premières valeurs du tableau la majorité des $n_{i,j}$ vaut 0, ce qui signifie que les moyennes réelles $h_p^{(i,j)}$ et $h_f^{(i,j)}$ sont en général calculées sur un seul prix (voir aucun). Par conséquent elles sont probablement beaucoup plus volatiles que si elles avaient été calculées sur 10 prix, il faudrait donc augmenter la valeur de σ pour reproduire ce phénomène.

A partir d'un niveau de 100, si on le perturbe multiplicativement par une loi log-normale de paramètres $\mu = 0$ et $\sigma = 0{,}05$ et que l'on simule un nombre suffisant de tirages, les valeurs obtenues sont dans leur grande majorité comprises entre 95 et 105. Si σ est fixée à 0.10, les résultats sont essentiellement distribués entre 85 et 115 ce qui pour des prix peut sembler raisonnable, compte tenu des différences de qualité. Pour approfondir l'étude des petits échantillons on fixe donc maintenant σ à 0.10 et $\mathcal{K}_0$ à 20 (les autres paramètres étant maintenus aux mêmes niveaux). Les lots de données comportent alors en moyenne 340 couples et les 2/3 des $n_{i,j}$ sont nuls. Les indicateurs de base donnent, pour 100 simulations, les résultats suivants :

Tableau 12 : Indicateurs d'erreur pour $\sigma = 0{,}10$ $K_0 = 20$

BIAIS	ERREUR MOYENNE	ECART-TYPE MOYEN	ERREUR_MAX MOYENNE	MAX ERREUR_MAX
0.16	5.54	3.39	14.17	25.46

L'erreur moyenne de 5% est raisonnable, mais l'aspect le plus problématique concerne les erreurs extrêmes. Pour un lot de données l'erreur maximale est en moyenne de 14.17% et lors des 100 simulations une de ces déviations maximales a atteint le niveau de 25.46%. Ces erreurs ne sont pas négligeables. Pour avoir une meilleure appréhension des déviations extrêmes les quantiles $Q_{x\%}$ sont calculés, en conservant l'hypothèse $\sigma = 0{,}10$ et en permettant à $\mathcal{K}_0$ de varier entre 5 et 50 (cf. tableau 13). La situation la moins défavorable se produit pour $\mathcal{K}_0 = 50$ (N = 860), mais même dans ce cas les résultats indiquent que l'indice ne peut pas être considéré comme entièrement fiable. Pour 10% de ses valeurs l'erreur est supérieure à 8% et la déviation maximale est de l'ordre de 15%.

Tableau 13 : Quantiles extrêmes pour $\sigma = 0{,}10$ en fonction de $\mathcal{K}_0$

$\mathcal{K}_0$	$Q_{10\%}$	$Q_{5\%}$	$Q_{1\%}$	N
5	22.03	27.32	39.59	86
10	14.90	17.97	24.70	172
20	11.07	13.19	17.30	342
30	9.28	11.07	14.62	516
40	8.60	10.51	13.74	688
50	8.21	9.82	13.40	860

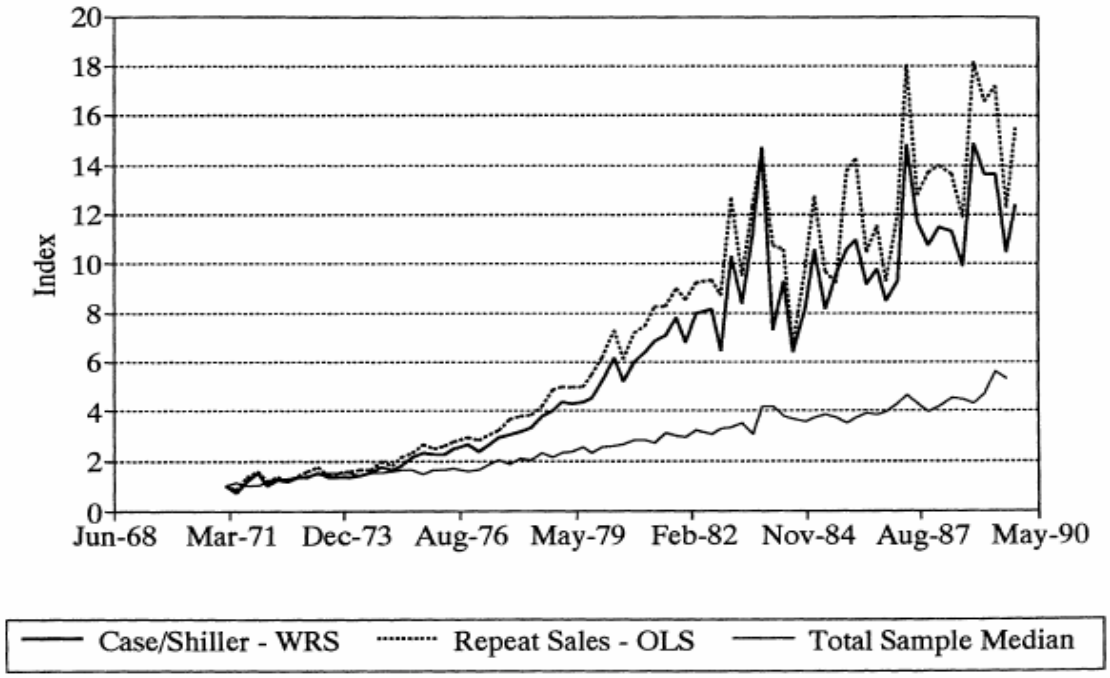
5.1.3. Comparaison avec un résultat empirique de la littérature

Ce problème de l'instabilité de l'indice dans les situations de données rares a déjà été rapporté dans la littérature. On peut se reporter par exemple à Gatzlaff, Geltner (1998) pour mesurer l'importance de cette question dans le domaine de l'immobilier commercial où les données sont en général beaucoup plus sporadiques que dans le secteur résidentiel[43]. L'article de Meese, Wallace (1997) fournit quant à lui les résultats obtenus pour la ville de Oakland lorsque l'indice est calculé pour 76 dates, à partir d'un échantillon d'environ 1700 ventes répétées. La figure 13 est directement extraite de cet article, elle présente le comportement de l'indice médian et des indices BMN (OLS : ordinary least squares) et Case-Shiller (WRS : weighted repeat sales) construits sur cet échantillon. A partir de 1979 l'ampleur des variations des indices de ventes répétées devient aberrante. Certains trimestres sont sujets à des hausses de 50% avant d'être suivis par des baisses tout aussi irréalistes. En appliquant le critère donnant la taille minimale requise pour que l'échantillon produise des indices stables, on obtient 5 x 75 x 76 = 28500 observations. Ce chiffre est très loin des 1700 données utilisées pour Oakland.

[43] Gatzlaff et Geltner proposent de traiter ce problème en appliquant la "méthode des moments". Cette technique vise à stabiliser l'indice en rajoutant une contrainte lors de son estimation. Dans l'article de 1998 la contrainte supplémentaire consiste à fixer le coefficient d'autocorrélation de l'indice à l'ordre 1 au niveau standard que l'on observe lorsque l'estimation est réalisée sur des échantillons de taille suffisante.

Si l'on veut reconstituer une situation similaire pour l'intervalle [0,40], la taille requise étant de 5 x 40 x 41 = 8200, il faut construire un échantillon comportant (1700/28500) x 8200 ≈ 490 couples. Le tableau 11 (ou 13) permet d'inférer[44] une valeur de 28 pour K_0. La figure 14 présente le type de graphique que l'on obtient alors pour [0,40], en gardant une volatilité σ à 0,10 et en fixant K_0 à 28.

Figure 13 : Indices obtenus par Messe, Wallace (1997) pour Oakland



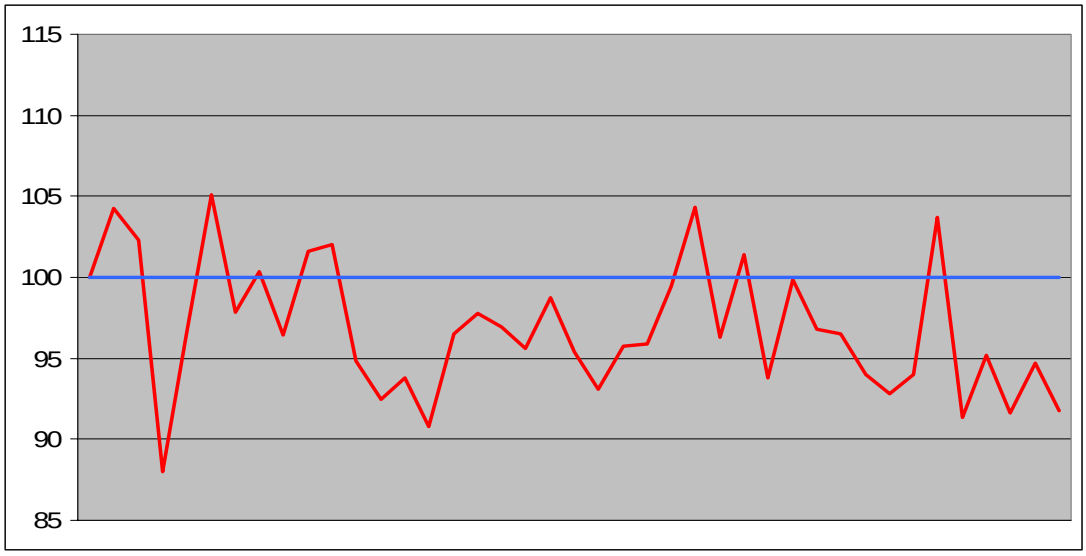
Les fluctuations ont des amplitudes comparables (figure 13 et figure 14), bien qu'elles soient un peu inférieures dans la figure 14. On peut cependant relever une différence importante. La première partie de l'indice d'Oakland (avant 1979) est relativement stable[45], tandis que dans la simulation il n'y a pas de période à faible volatilité. L'hypothèse naturelle que l'on peut avancer pour expliquer ce phénomène est l'hétérogénéité de la distribution temporelle des données. Si la période 1968-1979

[44] Par interpolation linéaire

[45] Même en tenant compte de l'effet de tassement de l'échelle, la volatilité dans la première partie est plus faible.

concentre par exemple 75% des données[46], les valeurs de l'indice pour cette première époque devraient être assez fiables. Corrélativement cela induirait alors une volatilité renforcée pour la deuxième époque qui est de longueur comparable. Les fluctuations de la figure 14 sont probablement représentatives de la volatilité moyenne de l'échantillon d'Oakland. Cette remarque nous amène maintenant à étudier les réactions de l'indice aux échantillons hétérogènes, et non plus seulement homogènes.

Figure 14 : RSI simulé dans des conditions équivalentes à celles de Messe, Wallace (1997)



Courbe bleue : « vrai prix » Courbe rouge : indice de ventes répétées

5.2. L'impact de l'hétérogénéité sur la fiabilité de l'indice

Dans cette section l'indicateur ERREUR_MOYENNE est décomposé pour pouvoir étudier la structure temporelle des erreurs. Les conséquences d'une réduction localisée de la densité des données sont ensuite analysées en détail.

[46] Cette remarque est d'ordre intuitif et elle ne se veut pas parfaitement rigoureuse car il faudrait alors aussi considérer les effets des couples à cheval sur ces deux périodes.

5.2.1. Désagrégation de l'erreur et effets de bords

L'indicateur ERREUR_MOYENNE agrège les pourcentages d'erreur entre toute les dates. Si on simule 100 jeux de données, la moyenne étant calculée sur 40 x 100 = 4000 observations, elle possède un niveau de stabilité satisfaisant. Pour l'étude de l'hétérogénéité il est nécessaire de désagréger cette grandeur en 40 sous-indicateurs ERREUR_MOYENNE(i) définis par :

$$\text{ERREUR\_MOYENNE}(i) = E\left[\,|e(i,\omega)|\,\right] \qquad i = 1, \ldots, 40$$

Ils représentent l'erreur moyenne, en valeur absolue, pour le i^{eme} intervalle [i – 1 , i]. Si l'on génère seulement 100 échantillons, chacune de ces 40 moyennes étant calculées sur 100 nombres, elles sont alors sujettes à des phénomènes de fluctuation non négligeables. Avoir des valeurs précises requerra souvent un minimum de 5000 simulations, le coût en termes de temps de calcul est donc d'un tout autre ordre lorsque l'on étudie les erreurs désagrégées.

La situation de référence pour ce paragraphe est définie par un choix de paramètres standard :

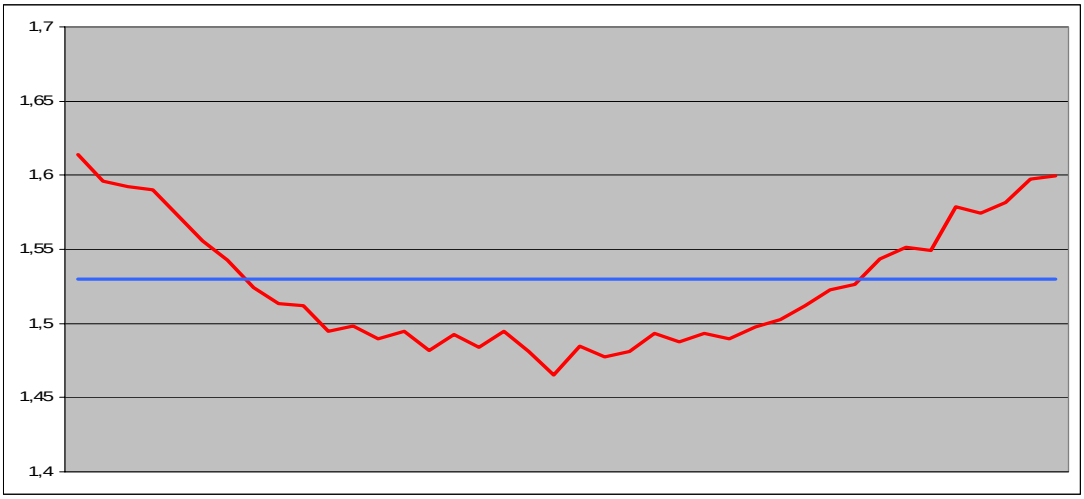
- une courbe des « vrais prix » plate (100 pour toutes les dates)
- un paramètre Θ fixé à 10
- une perturbation par une loi log-normale de paramètres $\mu = 0$ et $\sigma = 0.05$ pour les moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$
- un niveau de liquidité élevé et constant $K_i = 1$ et $\mathcal{K}_0 = 10000$.

L'étude de cette situation est réalisée en simulant 20000 échantillons. Le calcul de l'indicateur agrégé ERREUR_MOYENNE donne une valeur de 1.53, les valeurs désagrégées obtenues pour chacun des intervalles élémentaires sont présentées graphiquement dans la figure 15. On aurait pu penser a priori que les erreurs ne dépendraient pas de l'intervalle considéré, or ce n'est pas tout à fait le cas. Elles

forment une courbe symétrique en U, l'indice est donc plus précis au centre de l'intervalle que sur ses côtés. Cette particularité des erreurs n'est cependant pas d'une très grande ampleur puisque ERREUR_MOYENNE(0) ≈ 1,61 et ERREUR_MOYENNE(20) ≈ 1,47 ; les valeurs latérales ne sont supérieures que d'approximativement 10% à celles du centre pour ce choix de paramètres.

On retrouvera plus loin, dans le modèle des deux populations (chapitre 3, paragraphe 4.3), un phénomène similaire où le comportement d'une grandeur pour les premières et pour les dernières dates de l'intervalle, sera différent de celui observé au centre. L'indice de ventes répétées est en fait soumis à des phénomènes d'effets de bord. Si l'enjeu de cet artefact est faible pour l'extrémité gauche de l'intervalle, car elle est associée au passé lointain, il n'en va pas de même pour l'extrémité droite. Ce bord correspond en effet au passé très proche, le présent étant même souvent la dernière date de l'intervalle. Or, dans les études économiques ou pour les décisions d'investissement, ces dates sont les plus importantes car elles donnent par exemple des indications sur le momentum (souvent présent dans les courbes immobilières). Le phénomène des effets de bord doit donc être pris en considération convenablement.

Figure 15 : Indicateurs ERREUR_MOYENNE(i) pour la situation de référence



Courbe bleue : ERREUR_MOYENNE Courbe rouge : ERREUR_MOYENNE(i)

5.2.2. *Définition des chocs de liquidité*

Le choix de $\mathcal{K}_0 = 10000$ et $K_i = 1$ pour la situation de référence signifie que la liquidité du marché est bonne et que les échantillons sont suffisamment grands[47] pour considérer l'indice comme fiable. Etudier l'impact de l'hétérogénéité de la distribution n'est donc, a priori, pertinent que si les données se font plus rares pendant quelques périodes (on parlera de choc de liquidité), les effets de l'augmentation localisée de la densité des données ne devant pas avoir de conséquences significatives. On verra toutefois que, curieusement, cette intuition économétrique très classique ne se vérifiera pas systématiquement pour les indices de ventes répétées. L'augmentation du nombre des données pouvant, dans certains cas, engendrer une erreur plus importante.

L'intervalle choisi pour étudier les effets des chocs de liquidité est [18,22], afin d'éviter de possibles interférences avec les effets de bords. Un choc sera simulé en abaissant le paramètre de la loi binomiale de 0,03 à 0,0003 dans certaines zones de l'échantillon. Concrètement, si l'effectif encore en vie est de 10000 à la date t, en régime normal le nombre des reventes entre t et t + 1 sera en moyenne de 10000 x 0,03 = 300, contre seulement 10000 x 0,0003 = 3 lors des périodes d'assèchement de la liquidité.

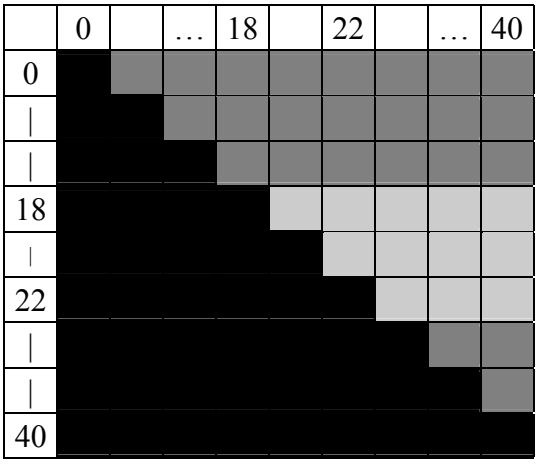
Une baisse du nombre de données utiles pour l'intervalle [18,22] peut se manifester de plusieurs manières. On parlera de choc « buy-side » (CBS) quand les ventes répétées dont la date d'achat se situe entre 18 et 22 seront sous-représentées dans l'échantillon. Symétriquement, un choc « sell-side » (CSS) se traduira par une sous représentation des données ayant une date de revente entre 18 et 22. Un choc mixte (CM) sera la conjonction de ces deux phénomènes.

[47] 170 000 ventes répétées en moyenne

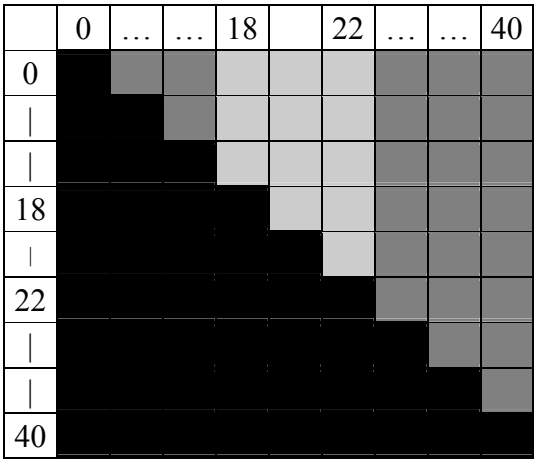
Pour générer un échantillon associé à un CBS il suffit de modifier le paramètre de la loi binomiale modélisant le devenir des cohortes[48] K_{18}, … , K_{22} en l'abaissant de 0.03 à 0.0003, jusqu'à la date t_{40}. Un CSS se simule en fixant le paramètre à 0,0003 pour les cohortes K_0, … , K_{21} lors de la simulation des reventes aux dates 18 à 22 (sur les dates antérieures et postérieures la valeur standard de 0,03 est conservée). Dans la distribution des $\{n_{i,j}\}$ on crée ainsi des zones à faible densité par rapport aux échantillons non choqués, cf. tableaux 14 et 15.

Tableau 14 : Densité des données pour un CBS et un CSS

Choc « buy-side » (CBS)



Choc « sell-side » (CSS)



Gris clair : densité faible Gris foncé : densité forte

Dans ces trois cas, le choc affecte les flux d'achat ou de vente entre 18 et 22 mais il ne modifie pas vraiment les effectifs pertinents pour ces intervalles. En effet, prenons l'exemple d'une transaction informationnelle pour l'intervalle [19,20], elle peut avoir sa date d'achat à t = 10 et sa date de revente à t = 30. La méthodologie appliquée pour générer les trois sortes d'échantillons choqués n'affecte donc a priori que faiblement les effectifs de ces types de données. Le quatrième choc consistera alors à réduire, pour les cohortes K_0,…,K_{21} , toutes les reventes se produisant à partir de[49] t_{18} en changeant la valeur du paramètre de la loi binomiale. Le tableau 15

[48] K_i désigne ici les ventes répétées initiées à la date "i" (achat à "i")

[49] et non plus seulement jusqu'à t_{22} comme dans le cas du CSS.

représente les densités obtenues dans cette situation, on parlera ici de choc sur les effectifs pertinents (CEP).

Tableau 15 : Densité des données pour un CM et un CEP

Choc mixte (CM)

	0		…	18		22		…	40
0									
|									
|									
18									
|									
22									
|									
|									
40									

Choc sur les effectifs pertinents (CEP)

	0	…	…	18		22	…	…	40
0									
|									
|									
18									
|									
22									
|									
|									
40									

Gris clair : densité faible Gris foncé : densité forte

5.2.3. Les effets des chocs

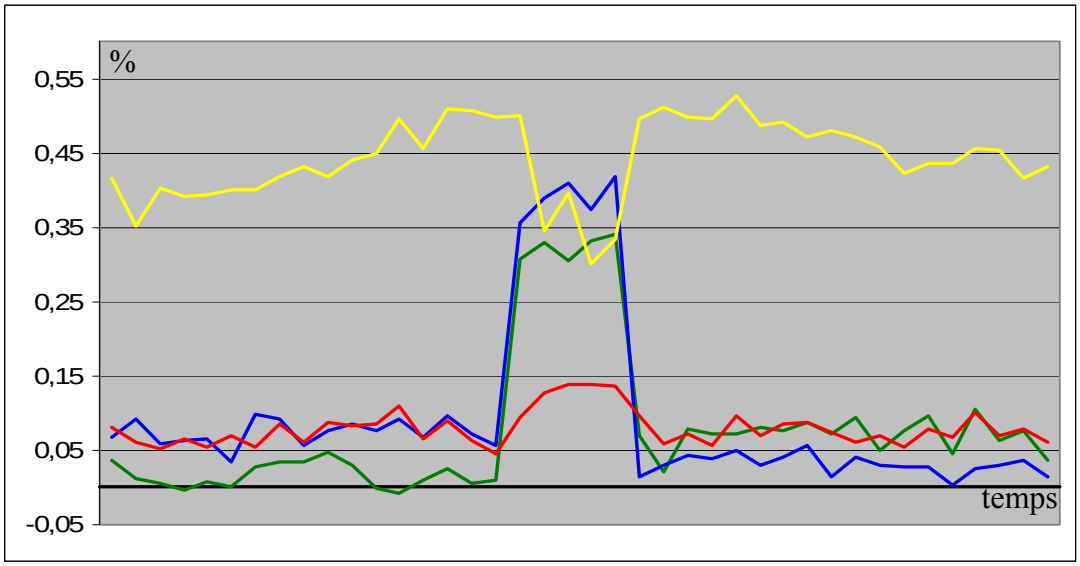
Pour chaque type de choc on génère 5000 échantillons. Les indicateurs ERREUR_MOYENNE(i) sont calculés et on leur soustrait les valeurs de l'échantillon standard[50]. La figure 16 illustre les surplus d'erreur obtenus selon les différents types de choc. En raison du plus faible nombre de simulations réalisées (5000 contre 20000) les résultats sont plus volatils que ceux de la figure 15, mais la précision est cependant suffisante pour analyser les tendances.

Pour les deux premiers cas (CBS et CSS), l'erreur observée sur l'intervalle [18,22] s'accroît significativement. Le saut est d'approximativement 0,35 pour un niveau de départ de l'échantillon benchmark à 1,47%, soit une augmentation de 25%.

[50] Retrancher directement l'indicateur agrégé ERREUR_MOYENNE = 1.53 à toutes les dates ne permettrait pas de supprimer les effets de bords ; l'interprétation serait alors biaisée.

On aurait pu penser a priori que le surplus d'erreur pour un choc mixte aurait été de 0,70 (surplus CBS + surplus CSS) or, comme on peut le voir sur la figure 16, il est beaucoup plus faible et n'atteint même pas 0,15. On détecte bien un léger accroissement pour les dates de 18 à 22 par rapport aux autres dates, mais ce phénomène est très ténu et la courbe rouge pourrait quasiment être considérée comme constante. On obtient donc ici un résultat inattendu : à partir d'un échantillon CBS, si l'on réduit le volume des données en simulant un CSS simultanément, l'erreur diminue ! L'intuition économétrique classique, qui nous amène à penser que moins les données sont nombreuses plus les erreurs seront importantes[51], ne semble pas devoir s'appliquer dans cette situation particulière.

Figure 16 : Surplus d'erreur pour les différents chocs de liquidité



Courbe verte : CBS Courbe bleue : CSS Courbe rouge : CM Courbe jaune : CEP

L'autre phénomène marquant concerne le CEP. La courbe se décale, dans son ensemble, nettement vers le haut, mais pour l'intervalle [18,22] on observe une moindre ampleur du phénomène (on passe de +0,50 à +0,35). Ce résultat est

[51] Par exemple, dans une régression par MCO, la variance des coefficients estimés décroît quand le nombre de données augmente.

également un peu surprenant car la motivation centrale qui a guidé la spécification du CEP était de réduire les données pertinentes pour [18,22]. On aurait donc pu espérer que le surplus d'erreur aurait été maximal sur cet intervalle, or c'est exactement le contraire qui se produit. La période [18,22] est la moins affectée par le choc …

Globalement, la réduction localisée de la densité des données engendre donc une plus grande imprécision car les valeurs associées aux quatre courbes sont toutes strictement positives. Toutefois, certains phénomènes contre-intuitifs viennent contredire cette affirmation, comme on a pu le constater.

5.3. Asymétrie des données et fiabilité de l'indice

5.3.1. Définition de l'asymétrie

L'hypothèse proposée pour expliquer les observations curieuses du paragraphe précédent est celle d'une sensibilité de l'indice à l'asymétrie de l'échantillon : si le "nombre" d'achats et le "nombre" de ventes à une date donnée dans un échantillon de ventes répétées ne sont pas « homogènes » l'erreur sera plus importante.

Ce concept d'homogénéité est assez ambigu, car par construction le nombre d'achats à t_2 est nettement supérieur au nombre de ventes à cette même date. La colonne 2 (vente à t_2) de la distribution des $\{n_{i,j}\}$ ne comporte par exemple que deux cases, alors que la ligne 2 (achat à t_2) en comporte 38. En fait, la définition de l'asymétrie ne se fera pas en termes d'effectif absolu, mais en termes de moyenne par classe de ventes répétées. On considérera ainsi que l'échantillon est homogène à la date 2 si le nombre moyen des effectifs par cellule dans la colonne 2 est proche de la moyenne des effectifs par cellule dans la ligne 2, et qu'il est asymétrique sinon.

5.3.2. Asymétrie et chocs de liquidité

Pour la date 20, un CBS appliqué à l'échantillon de référence (7.2.1) modifiera très fortement la moyenne de la ligne 20, mais n'affectera que faiblement celle de la colonne 20. L'asymétrie de l'échantillon est accentuée[52], la fiabilité se réduit. Si l'on y ajoute maintenant un CSS, on se retrouve dans le cas d'un CM. Les deux moyennes, ligne 20 et colonne 20, sont modifiées simultanément à la baisse mais leur niveau relatif est moins affecté que dans le cas d'un choc centré uniquement sur l'achat. L'asymétrie diminue, expliquant ainsi la curieuse réduction d'erreur observée pour le CM dans la figure 16.

Le cas du CEP vient confirmer l'hypothèse énoncée ci-dessus. Pour ce choc la moyenne des ventes à t_2 n'est pas modifiée, contrairement à celle des achats à t_2. L'asymétrie achat/vente est donc plus forte que pour l'échantillon non choqué. Au fur et à mesure que l'on se rapproche de la date 18, la moyenne des achats est de plus en plus réduite, tandis que celle des ventes est intacte. L'asymétrie augmente et cette situation correspond à la partie ascendante de la courbe jaune pour les dates de 0 à 18 (figure 16). Par contre entre 18 et 22 le choc va agir très symétriquement, réduisant ainsi les erreurs observées à ces dates. Elles restent cependant à un niveau plus élevé que pour l'échantillon de référence (+ 0,35), mais cela est probablement dû, cette fois, à l'effet des données rares.

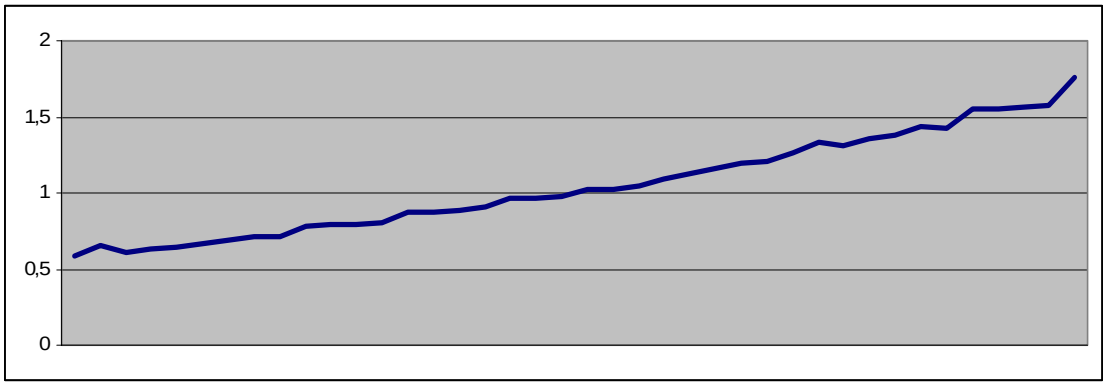
5.3.3. Asymétrie et effets de bords

La prise en compte de l'asymétrie pourrait également expliquer la forme en U observée pour les erreurs de l'échantillon de référence, cf. figure 15 paragraphe 5.2.1. Par construction, la simulation binomiale (ou exponentielle pour les lots de

[52] Pour le visualiser, on pourra se servir des Figures 15 et 16. Les zones en gris foncé ont les mêmes caractéristiques que l'échantillon non choqué ; par contre celles en gris clair ont une densité très réduite.

données suffisamment grands) génère des valeurs maximales pour les éléments $n_{t,t+1}$ situés sur la diagonale supérieure. Si l'on s'écarte de ces valeurs vers la droite ou vers le haut, les $n_{i,j}$ décroissent à des rythmes approximativement comparables. Pour la date t_2 la décroissance vers la droite à partir de $n_{1,2}$ pourra être longue, tandis qu'elle s'arrêtera très rapidement quand on se décalera vers le haut. Le nombre moyen des transactions par cellule est donc supérieur pour les reventes à t_2. D'une manière générale, au début de l'intervalle [0,T] la moyenne des effectifs par cellule à la revente est supérieure à son équivalent à l'achat. L'égalité se produit au milieu de l'intervalle et le rapport s'inverse ensuite. Le rapport de ces deux moyennes est représenté graphiquement sur la figure 17.

Figure 17 : Rapport entre le nombre moyen de transactions par cellule à l'achat et son équivalent à la vente, pour l'échantillon benchmark, sur l'intervalle [0,40]

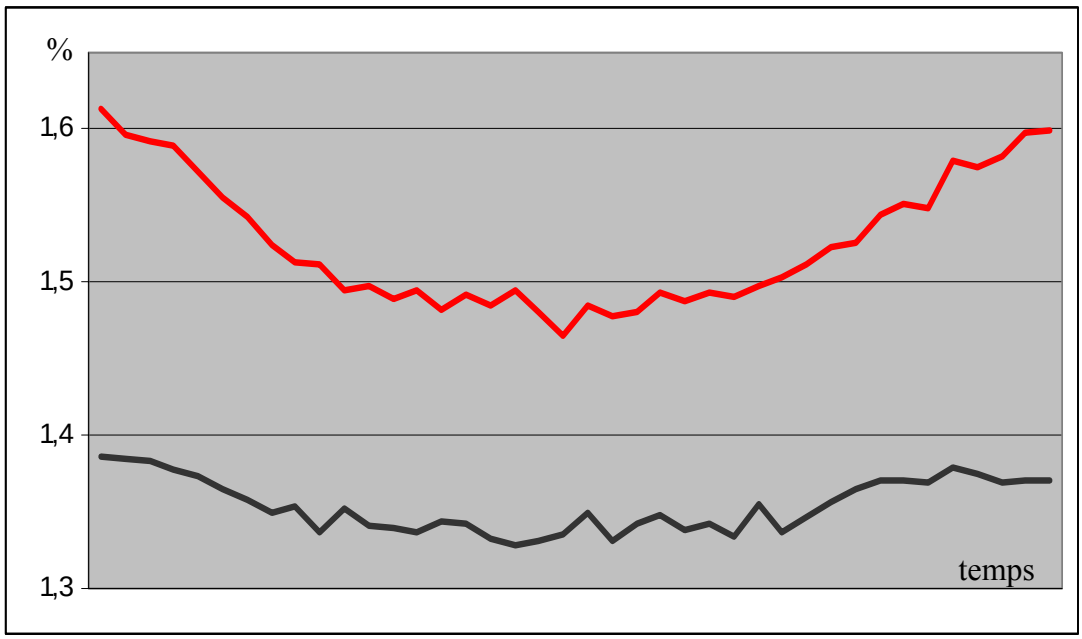


L'asymétrie de l'échantillon est donc plus marquée aux bords (ratio ≠ 1) et n'apparaît que très faiblement au centre (ratio ≈ 1). En appliquant l'hypothèse du 5.3.1 on doit par conséquent pouvoir observer des erreurs plus élevées aux bords : on retrouve évidemment ici la courbe en U de la figure 15.

Pour confirmer cette explication, on recalcule la courbe ERREUR_MOYENNE(i), mais en simulant cette fois un échantillon constant où les effectifs $n_{i,j}$ valent 100 pour tous les couples (i,j). Un tel échantillon ne présente pas d'asymétrie car les moyennes à l'achat et à la vente valent toutes 100. 20 000 lots de données sont simulés et les

résultats sont représentés dans la figure 18, la courbe en U de la figure 15 y est également reproduite à titre de comparaison.

Figure 18 : ERREUR_MOYENNE(i) pour le benchmark binomial et l'échantillon constant



Courbe rouge : échantillon binomial Courbe noire : échantillon constant

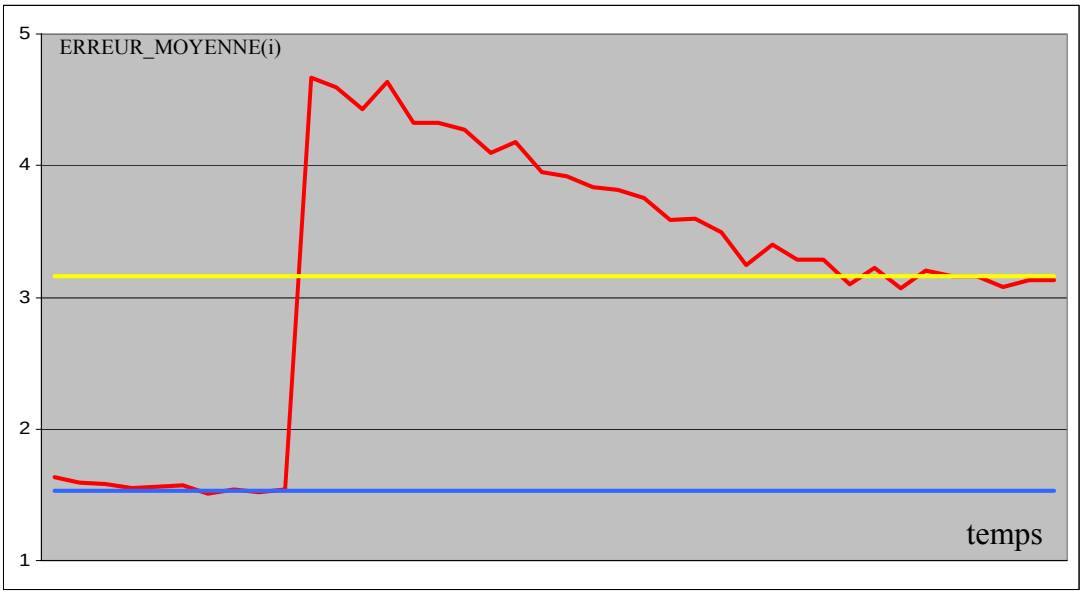
Si on constate que l'erreur dans le cas de l'échantillon constant semble augmenter légèrement aux bords de l'intervalle, la courbure reste cependant très faible. On peut d'ailleurs la considérer comme nulle car il est possible que d'autres phénomènes de bords, non liés à l'asymétrie, viennent encore perturber marginalement les résultats[53]. Globalement, le graphique confirme l'importance de l'asymétrie achat/vente pour les erreurs et vient étayer l'hypothèse selon laquelle la forme en U en est une conséquence.

[53] L'indice étant avant tout une grandeur informationnelle, un test avec un échantillon informationnellement constant ($L_{i,j}$ = 100) est sans doute plus légitime. On peut raisonnablement espérer que la légère courbure qui subsiste dans la figure 18 disparaîtrait alors complètement avec un tel choix.

5.3.4. Une conséquence inattendue de l'asymétrie

Dans ce contexte, il peut arriver que l'augmentation de la taille d'un échantillon réduise la fiabilité de l'indice. Ce phénomène se produit lorsque les nouvelles données sont très déséquilibrées entre l'achat et la vente. Le supplément d'information fourni par ces transactions complémentaires est alors surpassé par l'effet adverse de l'asymétrie, engendrant un niveau d'erreur supérieur. A titre d'exemple, la figure 19 représente le comportement de ERREUR_MOYENNE(i) lorsque l'on fait passer la cohorte K_{10} des biens négociés sur le marché à la date t = 10, d'un niveau de 10 000 (échantillon non choqué) à 1 000 000. Tous les autres paramètres sont conservés constants.

Figure 19 : Effet d'une augmentation asymétrique de la taille de l'échantillon sur l'erreur



Courbe bleue : ERREUR_MOYENNE avant modification
Courbe jaune : ERREUR_MOYENNE après modification
Courbe rouge : ERREUR_MOYENNE(i) pour l'échantillon modifié

Avant t_{10} l'asymétrie originelle de l'échantillon n'est pas modifiée, la courbe rouge et la courbe jaune sont donc au même niveau. A partir de t_{10} la modification de

la cohorte introduit une dissymétrie beaucoup plus forte dans l'échantillon. On constate alors que les erreurs augmentent très nettement ; l'effet est maximal à t_{10} mais il reste très marqué pour les dates postérieures. L'asymétrie est ici la principale source d'erreur. Elle fait passer du niveau standard ERREUR_MOYENNE = 1.53%, principalement causé par un phénomène de fluctuation d'échantillonnage, à des niveaux voisins de 4%. L'erreur a été multipliée par 2,5 !

5.4. Conclusion

La distribution de l'échantillon influence significativement, et de manières multiples, le degré de fiabilité de l'indice :

- Les lots de données trop réduits produisent des estimations volatiles. La taille critique permettant d'avoir une certaine confiance dans l'indice[54] est de 5T(T+1).
- La baisse localisée de la densité des données augmente en général le niveau d'erreur.
- L'asymétrie des échantillons agit négativement sur la fiabilité des indices. Elle produit dans certaines situations particulières des résultats surprenants venant contredire l'intuition classique, tel que l'augmentation des erreurs lorsqu'un échantillon est complété par de nouvelles données très asymétriques. En corollaire de ce résultat, on peut donc affirmer que dans certains cas d'échantillons fortement asymétriques, réduire correctement[55] le nombre des données revient à augmenter la fiabilité de l'indice.

[54] Sous les conditions de simulations présentées
[55] "correctement" doit s'entendre comme un rétablissement de la symétrie

6. Conclusion

En travaillant à partir d'un échantillon synthétique mais réaliste et en s'appuyant sur la réécriture du RSI exposé dans le chapitre précédent, cette section a mis en œuvre la méthodologie d'analyse des données de ventes répétées. Cette technique ne consiste plus à calculer uniquement des valeurs indicielles, on peut dorénavant profiter de la richesse de la structure théorique du RSI pour réaliser des études fines et détaillées.

Il a ensuite été établi que l'indice n'était pas sensible au contexte économique. Les effets du temps d'égalité des bruits et de la volatilité des prix sur la fiabilité ont également été étudiés empiriquement. Cette analyse a été menée dans une situation simplifiée, mais toutefois suffisamment précise pour avoir une appréhension correcte de l'influence de ces paramètres. Ces résultats pourront être approfondis ultérieurement.

Enfin, le dernier paragraphe a mis en évidence plusieurs phénomènes importants affectant la qualité de l'indice et qui n'ont encore jamais été évoqués dans la littérature, tels que les effets de bords ou le rôle de l'asymétrie. L'effet des chocs de liquidité a été mesuré et on a pu établir que l'augmentation du nombre de données pouvait, dans certains cas, avoir une influence négative sur la fiabilité.

Chapitre 3

Volatilité, réversibilité et problème des deux populations

1. Introduction

Le chapitre 2, en s'appuyant sur la reformulation théorique du chapitre 1 et sur le procédé synthétique de génération d'échantillons de ventes répétées, a mis en œuvre la méthodologie d'analyse de données et a étudié la fiabilité du RSI. Ces thèmes constituent un prolongement empirique direct des développements théoriques initiaux. Dans cette section, trois problématiques classiques de la littérature des ventes répétées seront étudiées à l'aune de ce nouveau formalisme ; on pourra ainsi juger de sa pertinence et de son efficacité pour résoudre des problèmes moins basiques que ceux du chapitre 2. On étudiera dans un premier temps la volatilité de l'indice, une formule très simple et très intuitive sera établie. Dans un second temps la question de la réversibilité des indices de ventes répétées sera analysée à l'aide de ces nouveaux concepts. On aboutira, ici aussi, à une formule claire et maniable permettant de quantifier facilement l'amplitude de ce phénomène indésirable. Pour résoudre ces deux problèmes les démarches seront similaires : on s'appuiera sur la décomposition de l'indice en briques élémentaires et, en partant des constituants les plus simples (les transactions), on remontera progressivement au niveau le plus élevé (l'indice). Enfin, la dernière partie de ce chapitre étudiera les conséquences de la présence de deux populations distinctes dans l'échantillon d'estimation. On verra notamment que, dans certains cas, l'indice ne retranscrira les caractéristiques que d'une seule des deux populations, l'autre étant muette économétriquement.

2. L'étude théorique de la volatilité

Ce paragraphe débute par l'exposé de quelques propriétés mathématiques de la matrice d'information. On précisera ensuite certaines des hypothèses du modèle de Case-Shiller. Ces deux premiers points permettront d'étudier rigoureusement et de manière approfondie la volatilité de l'indice. La formule finale sera simple et

intuitive, elle établira un lien entre la matrice de variance-covariance du vecteur R des taux de croissance de l'indice et la matrice d'information associée à l'échantillon.

2.1. Les propriétés mathématiques de la matrice d'information Î

La première proposition résume les propriétés les plus basiques de Î :

Proposition 1

- Pour $p \leq q$, le (p, q)-élément[1] $\hat{I}_{p,q}$ de Î, est $I^{[p-1,\, q]}$ et pour $p > q$, $I^{[q-1,\, p]}$
- Les valeurs sont toutes positives ($\hat{I}_{p,q} \geq 0$) et Î est symétrique ($\hat{I}_{p,q} = \hat{I}_{q,p}$)
- Les termes décroissent en ligne et en colonne à partir des éléments diagonaux

$$\hat{I}_{p,p} \geq \hat{I}_{p,p+1} \geq \ldots \geq \hat{I}_{p,T} \quad \text{et} \quad \hat{I}_{p,0} \leq \hat{I}_{p,1} \leq \ldots \leq \hat{I}_{p,p} \qquad \text{pour } p = 0, \ldots, T$$

$$\hat{I}_{p,p} \geq \hat{I}_{p+1,p} \geq \ldots \geq \hat{I}_{T,p} \quad \text{et} \quad \hat{I}_{0,p} \leq \hat{I}_{1,p} \leq \ldots \leq \hat{I}_{p,p} \qquad \text{pour } p = 0, \ldots, T$$

En utilisant la définition de $I^{[p-1,\, q]}$ et son interprétation comme somme partielle d'éléments du tableau des $\{L_{i,j}\}$, cf. tableau 1, on peut établir les relations suivantes pour la partie supérieure[2] de la matrice.

Proposition 2

- $\hat{I}_{p,q} = \hat{I}_{p-1,q} + \hat{I}_{p,q+1} - \hat{I}_{p-1,q+1} + L_{p-1,q}$ pour $1 < p \leq q < T$
- $\hat{I}_{1,q} = \hat{I}_{1,q+1} + L_{0,q}$ pour $1 \leq q < T$
- $\hat{I}_{p,T} = \hat{I}_{p-1,T} + L_{p-1,T}$ pour $1 < p \leq T$

$L_{i,j}$ étant toujours positif ou nul on obtient comme corollaire l'inégalité[3] :

- $\hat{I}_{p,q} \geq \hat{I}_{p-1,q} + \hat{I}_{p,q+1} - \hat{I}_{p-1,q+1}$ pour $1 < p \leq q < T$

[1] La matrice est indexée par $1 \leq p,q \leq T$

[2] Ces relations peuvent être facilement adaptées pour la partie inférieure de la matrice.

Tableau 1 : quantité d'information pertinente pour [p-1,q] : $I^{[p-1,q]} = I_{p,q}$

	0	…	p-1	…	q	…	T
0			$L_{0,\,p-1}$		$L_{0,\,q}$		$L_{0,\,T}$
⋮							
p-1					$L_{p-1,\,q}$		$L_{p-1,\,T}$
⋮							
T							

La trace[4] de la matrice $\hat{I}$, $I^0 + I^1 + \ldots + I^{T-1}$, peut se calculer à partir de la distribution informationnelle en remarquant que chaque $L_{i,j}$ apparaît exactement (j – i) fois dans cette somme :

$$Tr(\hat{I}) = \sum_{i<j}(j-i)L_{i,j} = \sum_{i<j}(j-i)\sum_{k,}(\Theta + (j-i))^{-1} = \sum_{i<j}\sum_{k,}G(j-i)$$

On a alors : $$Tr(\hat{I}) = N\,G(\zeta) \qquad (1)$$

D'autre part, si on introduit la somme des éléments juste au dessus de la diagonale principale $Tr_{+1}(\hat{I}) = I^{[0,2]} + I^{[1,3]} + \ldots + I^{[T-2,T]}$, on obtient la relation[5] :

$$Tr(\hat{I}) - Tr_{+1}(\hat{I}) = \sum_{t=0,\ldots,T-1}\sum_{i=0,\ldots,t} L_{i,t+1} = I \qquad (2)$$

Ces deux formules ont un intérêt pratique évident. Elles signifient, en effet, que l'on peut retrouver les mesures agrégées I (quantité d'information totale) et NG(ζ) (effectif total de l'échantillon, au coefficient G(ζ) près), par une simple lecture des éléments diagonaux de $\hat{I}$.

[3] Les inégalités pour $\hat{I}_{1,q}$ et $\hat{I}_{p,T}$ sont déjà connues
[4] La trace d'une matrice carrée désigne la somme de ses éléments diagonaux.
[5] cf. annexe 15

Comme toujours dans cette structure, s'il existe des relations au niveau agrégé il devrait également en exister pour les équivalents variables temporellement, c'est-à-dire pour I^t et n^t. On devrait donc pouvoir retrouver ces deux séries de mesures directement grâce à la matrice Î. Pour la première grandeur, la formule consiste simplement à remarquer que ces nombres correspondent aux éléments diagonaux de Î. Pour n^t la formule est moins triviale (cf. annexe 15), elle s'écrit:

$$\sum_{t' \leq t} I^{[t', t+1]} + \sum_{t' > t} I^{[t, t'+1]} = n^t \, G(\zeta^t) \qquad (3)$$

Concrètement, la somme des éléments d'une ligne de Î fournit la valeur n^t correspondante, au coefficient $G(\zeta^t)$ près. La proposition ci-dessous récapitule les formules 1, 2 et 3.

Proposition 3

- $Tr(\hat{I}) = N \, G(\zeta)$
- $Tr(\hat{I}) - Tr_{+1}(\hat{I}) = I$
- $\sum_{t' \leq t} I^{[t', t+1]} + \sum_{t' > t} I^{[t, t'+1]} = n^t \, G(\zeta^t)$

2.2. Les hypothèses d'indépendance

On précisera et on complètera dans ce paragraphe les hypothèses portant sur les processus intervenant dans la modélisation, afin de pouvoir étudier la volatilité de l'indice et de ses divers constituants. Le principe de la démonstration consistera à remonter des transactions à l'indice, en passant par les grandeurs intermédiaires mises en évidence lors de la décomposition théorique du RSI.

2.2.1. Description de l'indice

Dans le calcul d'un indice de Case-Shiller le prix d'un bien immobilier k, à une date donnée t, est décomposé en trois éléments : la composante commune à toutes les transactions réalisées à cette date (l'indice), les tendances spécifiques et le bruit résultant des imperfections du marché. De manière formelle, cette relation s'écrit :

$$\ln(p_{k,t}) = \ln(\text{Indice}_t) + G_{k,t} + N_{k,t} \qquad (4)$$

en notant : - $p_{k,t}$ le prix du $k^{ème}$ bien à la date t
- indice_t la valeur théorique de l'indice à t

Pour chaque bien k, la déviation du prix par rapport à l'indice est modélisée à l'aide de deux processus stochastiques à temps discret, $(G_k)_{t= 0,\ldots,T}$ et $(N_k)_{t=0,\ldots,T}$. Les valeurs de l'indice sont également aléatoires mais la source de cette incertitude est d'une nature différente ; il s'agit d'un aléa économique et non pas d'un aléa d'échantillonnage comme pour G_k et N_k. On supposera que le contexte économique est donné et qu'il n'affecte pas les lois d'échantillonnage, ou en d'autres termes, que les valeurs Indice_t sont des constantes et non pas des variables aléatoires. Quand on parlera de la distribution de $p_{k,t}$, il s'agira en fait plus exactement de la loi de $p_{k,t}$ conditionnellement à la valeur Indice_t. De même les dynamiques des différentes grandeurs étudiées ci-après seront implicitement des dynamiques conditionnelles[6].

2.2.2. La modélisation des tendances spécifiques

Le premier terme de la déviation modélise pour chaque bien sa tendance spécifique par une marche aléatoire gaussienne G_k.

[6] C'est-à-dire sachant les valeurs Indice_t

Ainsi :
- $G_{k,t}$ suit une loi normale $\mathcal{N}(0 ; \sigma_G^2 t)$
- Le processus est à accroissements indépendants
- On notera par $G_{k,0}$ son point de départ

On supposera de plus que :

(H1) Pour un nombre suffisamment important de propriétés hétérogènes la moyenne des valeurs initiales $G_{k,0}$ est nulle.

(H2) Les processus stochastiques (G_k) sont indépendants dans leur ensemble.

Dans les articles classiques ces deux hypothèses mineures ne sont traditionnellement pas évoquées mais, en toute rigueur, elles sont nécessaires pour arriver à des formules simples. La première se comprend en formulant l'hypothèse inverse, c'est-à-dire en supposant que la moyenne des points de départ des tendances spécifiques est sensiblement non nulle. Les biens étant suffisamment nombreux et diversifiés, cet écart de valeur s'appliquerait alors aussi au marché dans son ensemble. Par conséquent il devrait être capturé par $Indice_0$, et non pas par les $G_{k,t}$. Une moyenne non nulle traduit donc l'existence d'un comportement systématique, non pris en compte par l'indice. Cette situation est contradictoire et justifie l'hypothèse (H1).
La deuxième hypothèse n'est pas énoncée explicitement dans les articles de référence de la littérature, mais elle existe toutefois implicitement dans ces modèles. L'estimation de l'indice se fait à partir de la relation :

$$\text{Ln}(p_{k,j} / p_{k,i}) = \ln(\text{indice}_j / \text{indice}_i) + \underbrace{(G_{k,j} - G_{k,i}) + (N_{k,j} - N_{k,i})}_{\varepsilon_k} \qquad (5)$$

où l'on suppose que la matrice de variance-covariance des résidus est diagonale, il n'y a donc pas d'autocorrélation des ε_k. Or, pour assurer cette propriété la condition d'indépendance des processus (G_k) est une condition nécessaire[7], l'hypothèse (H2) est donc requise. D'un point de vue formel elle ne pose pas de réelles difficultés, mais on peut néanmoins s'interroger sur son degré de vraisemblance économique. Elle affirme que les composantes spécifiques à chaque bien sont indépendantes. Cependant, si l'on considère deux biens voisins géographiquement et assez similaires cela peut sembler discutable : leurs évolutions idiosyncratiques sont probablement assez corrélées. Dans le cas de deux propriétés moins assimilables l'hypothèse d'indépendance est raisonnable, mais affirmer qu'il en est toujours ainsi est sans doute légèrement inexact. Le but poursuivi ici n'étant pas de construire un indice prenant en compte ces petites autocorrélations potentielles, on supposera donc l'hypothèse (H2) satisfaite, comme cela se fait classiquement.

2.2.3. La modélisation des imperfections du marché

Le troisième terme dans la décomposition du prix d'un bien k est le bruit blanc (N_k), que l'on supposera gaussien :

- $N_{k,t}$ suit une loi normale $\mathcal{N}(0 ; \sigma_N^2)$
- Pour un même bien k, les variables aléatoires $N_{k,t}$ et $N_{k,t'}$ sont indépendantes ($t \neq t'$)

Usuellement, la nature du bruit blanc n'est pas précisée, mais pour pouvoir travailler explicitement avec les lois des différentes composantes de l'indice il est nécessaire de la spécifier. On choisira donc un bruit blanc gaussien.

[7] On assimile ici abusivement les notions de non-corrélation et d'indépendance dans un souci de simplicité.

Les imperfections seront de plus indépendantes des tendances idiosyncratiques :

(H3) Les processus $(N_k)_{k=0,\ldots,N}$ sont indépendants des processus $(G_k)_{k=0,\ldots,N}$

2.3. Niveau de la vente répétée

2.3.1. La dynamique des prix

Pour un bien k et à une date t, la déviation du prix par rapport à l'indice[8] est égale à $G_{k,t} + N_{k,t}$. En utilisant l'hypothèse d'indépendance (H3) on en déduit immédiatement que cette somme suit une loi normale d'espérance $G_{k,0}$ et de variance $\sigma_N^2 + \sigma_G^2\, t = \sigma_G^2\,(\Theta / 2 + t)$.

La relation (4) peut se réécrire sous la forme :

$$p_{k,t} = \text{indice}_t \ \exp(G_{k,t} + W_{k,t}) = \text{indice}_t \ C(p_{k,t}) \qquad (6)$$

Cette modélisation fait donc l'hypothèse qu'un prix se déduit de l'indice par un coefficient multiplicatif $C(p_{k,t})$ suivant une loi log-normale $\mathcal{LN}\left(G_{k,0}\,;\, \sigma_G^2\,(\Theta/2 + t)\right)$

2.3.2. La dynamique des rendements

Le bien k est acheté à t_i, revendu à t_j, son taux de rentabilité est $r_k^{(i,j)} = \ln(p_{k,j}/p_{k,i})/(j-i)$.
Ce taux peut se réécrire en décomposant les prix en leurs trois composantes :

$$r_k^{(i,j)} = \left[\ \ln(\ \text{Indice}_j / \text{Indice}_i\ ) + (\ G_{k,j} - G_{k,i}\ ) + (\ N_{k,j} - N_{k,i}\ )\ \right] / (\ j - i\ ) \qquad (7)$$

[8] en logarithme

Les accroissements de la marche aléatoire et du bruit blanc gaussien suivent des lois normales, respectivement $\mathcal{N}(0;\sigma_G^2\,(j-i))$ et $\mathcal{N}(0;2\sigma_N^2)$. En utilisant l'hypothèse (H3) d'indépendance des deux bruits on peut en déduire la loi suivie par $r_k^{(i,j)}$. Il s'agit d'une loi normale de paramètres $\mathcal{N}(\ln(Indice_j / Indice_i) / (j-i) ; (\sigma_G^2(j-i)+2\sigma_N^2)/(j-i)^2)$

En notant $r^{(i,j)} = \ln(Indice_j / Indice_i) / (j - i)$ et en remarquant que la variance peut se réécrire à l'aide de la fonction G comme $\sigma_G^2 / ((j - i) G(j - i))$, on obtient plus simplement :

$$r_k^{(i,j)} \sim \mathcal{N}(r^{(i,j)} ; \sigma_G^2 / ((j - i) G(j - i))) \qquad (8)$$

2.4. Niveau de la classe de ventes répétées (i,j)

2.4.1. La dynamique des prix

A partir des définitions des moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$ on peut écrire :

$$h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}} = indice_i \exp[(\Sigma_{k'} G_{k',i} + N_{k',i})/ n_{i,j}] = indice_i\, C(h_p^{(i,j)}) \qquad (9)$$

$$h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}} = indice_j \exp[(\Sigma_{k'} G_{k',j} + N_{k',j})/ n_{i,j}] = indice_j\, C(h_f^{(i,j)}) \qquad (9')$$

Des hypothèses d'indépendance (H2) et (H3) on déduit que les lois suivies par les coefficients $C(h_p^{(i,j)})$ et $C(h_f^{(i,j)})$ sont de type :

$\mathcal{LN}((\Sigma_{k'}G_{k',0}) / n_{i,j} ; \sigma_G^2\,(\Theta/2 + i)/n_{i,j})$ et $\mathcal{LN}((\Sigma_{k'}G_{k',0}) / n_{i,j} ; \sigma_G^2\,(\Theta/2 + j)/n_{i,j})$

Si on utilise maintenant pour les espérances l'hypothèse (H1) à l'intérieur de la classe[9] (i,j), les distributions sont alors de type :

$$\mathcal{LN}\left(0 ; \sigma_G^2(\Theta/2 + i) / n_{i,j}\right) \qquad \text{et} \qquad \mathcal{LN}\left(0 ; \sigma_G^2(\Theta/2 + j) / n_{i,j}\right) \qquad (10)$$

Il faut remarquer ici que la connaissance explicite de ces lois permettrait d'affiner le processus de génération des données synthétiques. En effet, dans le chapitre 2 paragraphe 2.6, les valeurs des moyennes $h_p^{(i,j)}$ et $h_f^{(i,j)}$ ont été déduites de la courbe des « vrais prix » en perturbant celle-ci par des coefficients multiplicatifs suivants des lois log-normale $\mathcal{LN}(0 ; 0{,}05)$. Or, si l'on accepte les hypothèses H1, H2 et H3, le second paramètre de cette loi ne peut pas être considéré comme constant. Une telle modification aurait sans doute une certaine influence sur le niveau des résultats obtenus mais elle ne causerait probablement pas de changements radicaux tels que la suppression des effets de bords ou la neutralisation du rôle de l'asymétrie. Le paragraphe 4 du chapitre précédent, qui étudiait les sensibilités de l'indice à divers paramètres, est vraisemblablement le lieu où cette modification serait la plus significative (en particulier dans la section 4.4 pour la sensibilité à la volatilité). La simplification consistant à utiliser une loi $\mathcal{LN}(0 ; 0{,}05)$ nous a cependant permis de nous faire une idée assez raisonnable des réactions de l'indice (en première approximation).

2.4.2. La dynamique des rendements

La rentabilité moyenne dans la classe (i,j) est un concept qui n'a pas été défini jusqu'à maintenant, on le notera par $\rho_{i,j}$. A l'intérieur de cette classe les effets des bruits sont identiques. Il n'est donc pas nécessaire d'introduire des pondérations

[9] On suppose donc que $n_{i,j}$ n'est pas trop petit et que les biens dans cette classe sont relativement diversifiés.

informationnelles pour définir la rentabilité moyenne qui se calculera simplement par :

$$\rho_{i,j} = (\sum_{k'} r_{k'}^{(i,j)}) / n_{i,j} = \ln(h_f^{(i,j)} / h_p^{(i,j)}) / (j - i) \qquad (11)$$

Les hypothèses (H2) et (H3) permettent d'affirmer que $\rho_{i,j}$ est une somme de variables aléatoires indépendantes, on obtient donc immédiatement sa loi :

$$\rho_{i,j} \sim \mathcal{N}(r^{(i,j)} ; \sigma_G^2 / (L_{i,j} (j - i)^2)) \qquad (12)$$

2.5. Niveau des effectifs pertinents pour [t, t+1]

2.5.1. La dynamique des prix

Le pas suivant dans l'étude des volatilités consiste à déterminer les lois suivies par $H_p(t)$ et $H_f(t)$. Ces moyennes sont définies par :

$$H_p(t) = \left(\prod_{i \le t < j} (h_p^{(i,j)})^{L_{i,j}} \right)^{1/I^t} \qquad H_f(t) = \left(\prod_{i \le t < j} (h_f^{(i,j)})^{L_{i,j}} \right)^{1/I^t} \qquad (13)$$

Remplaçons dans l'expression de $H_p(t)$, la quantité $h_p^{(i,j)}$ par le produit d'une constante et d'une variable aléatoire, $indice_i$ * $C(h_p^{(i,j)})$, comme indiqué dans le 2.4.1. Il vient alors :

$$H_p(t) = [\prod_{i \le t < j} (indice_i \, C(h_p^{(i,j)}))^{L_{i,j}}]^{1/I^t} = [\prod_{i \le t < j} indice_i^{L_{i,j}}]^{1/I^t} [\prod_{i \le t < j} C(h_p^{(i,j)})^{L_{i,j}}]^{1/I^t} \qquad (14)$$

Cette écriture présente le grand avantage de permettre le réemploi des calculs du paragraphe 3 (chapitre 1) pour le premier crochet, car on se retrouve dans la situation

simplifiée. Tous se passe, en effet, comme si les transactions étaient réalisées au niveau exact de indice_i (le indice_i utilisé ici correspond au h_i du paragraphe 3, chapitre 1). On sait donc que ce premier terme est une constante égale à la moyenne géométrique des valeurs passées[10], pondérées par les niveaux d'activité informationnelle côté achat : $\left[\prod_{i \leq t} \text{indice}_i^{B_i^t}\right]^{1/I^t}$.

Dans le deuxième crochet, les variables aléatoires $C(h_p^{(i,j)})^{L_{i,j}}$ suivent des lois log-normales de paramètres 0 et $(L_{i,j})^2 \sigma_G^2 (\Theta/2 + i) / n_{i,j}$. La quantité d'information $L_{i,j}$ a été définie par $L_{i,j} = n_{i,j} / (\Theta + (j - i))$, en redivisant par $\Theta + (j - i)$ on introduit une nouvelle grandeur :

$$V_{i,j} = L_{i,j} / (\Theta + (j - i)) = n_{i,j} / (\Theta + (j - i))^2 \qquad (15)$$

qui permet de réécrire le deuxième paramètre de la loi sous la forme $V_{i,j} \sigma_G^2 (\Theta/2+ i)$. Les hypothèses d'indépendance (H2) et (H3) permettent d'affirmer que ce deuxième crochet, que l'on notera $C(H_p(t))$, suit aussi une loi log-normale.

Le résultat final pour $H_p(t)$ consiste donc à écrire cette grandeur :

$$H_p(t) = \left[\prod_{i \leq t} \text{indice}_i^{B_i^t}\right]^{1/I^t} C(H_p(t)) \qquad (16)$$

- le premier terme est une constante qui représente le prix moyen d'achat, pondéré par l'information côté achat, quand les transactions se font aux niveaux $\{\text{indice}_t\}_{t=0,\ldots,T-1}$; c'est le $H_p(t)$ de la situation simplifiée
- la variable aléatoire $C(H_p(t))$ suit une loi $\mathcal{LN}\left(0 ; \sigma_G^2 (1/I^t)^2 \Sigma_{i \leq t < j} V_{i,j}(\Theta/2+ i)\right)$

[10] Valeurs passées de l'indice de prix dans la situation simplifiée

De manière analogue, $H_f(t)$ se décompose en :

$$H_f(t) = \left[\prod_{j>t} \text{indice}_j^{S_j^t}\right]^{1/I^t} C(H_f(t)) \qquad (16')$$

- le premier terme est le $H_f(t)$ de la situation simplifiée
- la variable aléatoire $C(H_f(t))$ suit une loi $\mathcal{LN}\left(0 ; \sigma_G^2 (1/I^t)^2 \Sigma_{i \le t < j} V_{i,j}(\Theta/2+ j)\right)$

La situation complexe se déduit donc directement de la situation simplifiée, par l'intermédiaire d'un coefficient multiplicatif dont on connaît la loi. L'intérêt d'avoir exploré préalablement le cas simplifié apparaît ici clairement.

2.5.2. La dynamique des rendements

On peut faire apparaître la moyenne des rentabilités[11] pour une classe (i,j) dans la formule de la moyenne des taux moyens ρ_t en écrivant :

$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j - i)\, n_{i,j}\ (\sum_{k'} r_{k'}^{(i,j)} / n_{i,j})$$

$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j - i)\, n_{i,j}\ \rho_{i,j} \qquad (17)$$

Grâce aux hypothèses (H2) et (H3) on peut alors en déduire la loi suivie par ρ_t :

$$\rho_t \sim \mathcal{N}\left((n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j - i)\, n_{i,j}\, r^{(i,j)} ; \sigma^2_G / ((\tau^t)^2 I^t)\right) \qquad (18)$$

[11] cf. paragraphe 2.4.2

L'espérance correspond à la valeur que l'on obtient pour ρ_t si l'on suppose que toutes les transactions se font aux niveaux des valeurs $\{indice_t\}_{t=0,\ldots,T}$. ρ_t est donc centré sur sa valeur du cas simplifié.

La formule de la variance est démontrée dans l'annexe 16. En écrivant que pour t, t' on a :

$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j - i) n_{i,j} \rho_{i,j} \quad \text{et} \quad \rho_{t'} = (n^{t'} G(\zeta^{t'}))^{-1} \sum_{i \le t' < j} G(j - i) n_{i,j} \rho_{i,j} \qquad (19)$$

on établit également dans cette même annexe la formule des covariances :

$$Cov(\rho_t ; \rho_{t'}) = \sigma_G^2 \, I^{[t', t+1]} / (\tau^t \tau^{t'} I^t I^{t'}) \qquad (20)$$

2.6. Niveau de l'indice estimé

La dynamique de l'indice ne peut pas se déduire des lois de $H_p(t)$ et $H_f(t)$, comme on aurait pu le penser en examinant la démarche suivie jusqu'ici. On l'obtiendra en fait via la dynamique des taux R.

2.6.1. La dynamique du vecteur des taux de croissance R

Etudions tout d'abord le vecteur η P. Les coordonnées de son vecteur des espérances sont :

$$\sum_{i \le t < j} G(j - i) \, n_{i,j} \, r^{(i,j)} \qquad (t = 0,\ldots,T-1)$$

Sa matrice de variance-covariance se déduit directement des résultats établis précédemment pour P.

On a ainsi : $$\mathrm{Cov}\left((\eta P)_t \, ; \, (\eta P)_{t'} \right) = \sigma_G^2 \, I^{[\,t',\, t+1]} \qquad (21)$$

avec le cas particulier : $$\mathrm{V}\left((\eta P)_t \right) = \sigma_G^2 \, I^{\,t} \qquad (22)$$

La matrice en question n'est donc rien d'autre que $\sigma_G^2 \, \hat{I}$.

Pour en déduire les caractéristiques aléatoires du vecteur R il suffit maintenant d'appliquer la formule fondamentale de l'indice de ventes répétées : $R = \hat{I}^{-1} (\eta \, P)$. Toute combinaison linéaire Λ des coordonnées de R est une combinaison linéaire des cordonnées de P et donc une combinaison linéaire des variables aléatoires gaussiennes indépendantes $r_{k'}^{(i,j)}$. Λ suit donc une loi normale et le vecteur R est alors un vecteur gaussien.

Le vecteur des espérances, E(R), est égal à $\hat{I}^{-1}$ E(ηP). Or comme il a déjà été mentionné, ce produit matriciel revient à calculer l'indice de ventes répétées en supposant que toutes les transactions se font aux niveaux $\{\text{indice}_t\}_{t=0,\ldots,T}$. Cette situation correspondant presque[12] exactement à celle du paragraphe 3 (chapitre 1), on peut alors affirmer que le résultat obtenu n'est rien d'autre que le vecteur des taux de croissance de l'indice théorique : $r'_t = \ln(\text{Indice}_{t+1} / \text{Indice}_t)$. Le vecteur R est donc un estimateur sans biais du vecteur des taux de croissance de l'indice théorique R*.

[12] Dans le paragraphe 3 du chapitre 1, on supposait que le bruit n'avait qu'une seule source (les tendances idiosyncratiques décrites par la marche aléatoire gaussienne G) or dans le cas qui nous occupe ici elles sont au nombre de deux, le paramètre Θ représentant la contribution de la deuxième source (les imperfections du marché décrites par le bruit blanc N). Toutefois si les transactions se font toujours au niveau de Ind, le résultat $r_t = \text{Ind}_{t+1} / \text{Ind}_t$ se maintient comme on peut s'en convaincre facilement en réexaminant le programme d'optimisation.

Si deux vecteurs aléatoires sont liés par une relation linéaire Y = BX, on sait que les matrices de variance-covariance le sont par la relation $\mathcal{V}Y = B\,(\mathcal{V}X)\,B'$. La matrice associée au vecteur R est alors :

$$\mathcal{V}(R) = \hat{I}^{-1}\,(\,\sigma_G^2\,\hat{I}\,)\,(\,\hat{I}^{-1})'$$

En simplifiant, et remarquant que si $\hat{I}$ est symétrique son inverse l'est aussi, on obtient :

$$\mathcal{V}(R) = \sigma_G^2\,\hat{I}^{-1} \qquad (23)$$

Le résultat final est donc particulièrement simple et intuitif. La matrice de variance-covariance de R est égale à l'inverse de la matrice d'information, au coefficient σ_G^2 près. Si on assimile, un peu abusivement, la matrice $\hat{I}$ à un nombre représentant la quantité d'information et si l'on considère $\mathcal{V}(R)$ comme un réel mesurant la taille moyenne de l'erreur, cette formule s'énonce alors synthétiquement :

« L'erreur est égale à l'inverse de l'information »

Ce résultat doit être mis en perspective avec les concepts introduits au début de la modélisation. La relation de départ[13] pour une transaction de la classe (i,j) était :

$$\begin{aligned} Ln(\,p_{k,j} / p_{k,i}\,) &= \ln(\,indice_j / indice_i\,) + (\,G_{k,j} - G_{k,i}\,) + (\,N_{k,j} - N_{k,i}\,) \\ &= \ln(\,indice_j / indice_i\,) + \varepsilon \end{aligned}$$

Assimiler le rendement de l'indice au rendement du bien (ou l'inverse) revient à faire une erreur de ε. La taille moyenne de cette erreur se mesure par la variance du résidu

[13] Formule (5)

$2\sigma_N^2 + \sigma_G^2(j-i)$. On a d'autre part défini la quantité d'information fournie par un bien particulier par $1/(\Theta+(j-i))$, où $\Theta = 2\sigma_N^2 / \sigma_G^2$. On a donc pour chaque vente répétée :

$$2\,\sigma_N^2 + \sigma_G^2\,(\,j - i\,) = \sigma_G^2 * (\,\Theta + (\,j - i\,)\,)$$

Soit : *Erreur potentielle d'une vente répétée* $= \sigma_G^2$ */(quantité d'information fournie)*

La relation $\mathcal{V}(R) = \sigma_G^2\ \hat{I}^{-1}$ n'est donc rien d'autre qu'une version agrégée de ce rapport entre erreur et information. Les concepts coïncident parfaitement entre le niveau indiciel et le niveau élémentaire des biens.

2.6.2. L'espérance de Lind

Dans le chapitre 1, paragraphe 2.1, le vecteur LIndice désignait le vecteur des logarithmes de l'indice théorique. Le vecteur Lind regroupait, quant à lui, les logarithmes de l'indice estimé. Lind et R sont liés par la matrice triangulaire A, dont les valeurs sont égales à 1 sur la diagonale principale et en dessous, 0 ailleurs. Plus précisement on a LInd = A R, comme R est un vecteur gaussien il en sera alors de même pour LInd.

L'espérance de LInd est donnée par le produit A E(R). Or, comme les coordonnées de E(R) sont de la forme $r'_t = \ln(\text{indice}_{t+1}/\text{indice}_t)$, les coordonnées du vecteur E(LInd) seront égales[14] à $\ln(\text{indice}_t)$. LInd est donc un estimateur sans biais du logarithme de l'indice théorique Lindice.

[14] Il suffit d'écrire le détail du produit matriciel pour s'en convaincre.

2.6.3. La matrice de variance-covariance de LInd

La matrice de variance-covariance de LInd est, en appliquant la formule énoncée ci-dessus (Y = BX => $\mathcal{V}$Y = B ($\mathcal{V}$X) B'), et en remarquant que A est inversible :

$$\mathcal{V}(\text{LInd}) = A\,\sigma_G^2\,\hat{I}^{-1}\,A' = \sigma_G^2\,A\,\hat{I}^{-1}\,A' = \sigma_G^2\,(\,A'^{-1}\,\hat{I}\;A^{-1}\,)^{-1} \qquad (24)$$

Ce résultat peut se réécrire sous une forme matricielle plus simple en introduisant deux nouvelles matrices de dimension T, $\mathcal{L}$ et $\mathcal{T}$ (cf. annexe 17 pour la démonstration). La matrice $\mathcal{L}$ n'est rien d'autre que le tableau des $L_{i,j}$ duquel la première ligne a été ôtée, complété par des zéros sur la diagonale et sur la partie triangulaire inférieure. La matrice $\mathcal{T}$ est une matrice diagonale donnant la quantité d'information fournie par les transactions de la date p (p = 1,…,T), sans tenir compte du type de la transaction. Par exemple pour p = 2, B^2_2 désigne la quantité d'information fournie par les ventes répétées de l'échantillon pour lesquelles l'achat a été réalisé à t = 2, et S^1_2 est son équivalent pour les reventes à t = 2.

$$\mathcal{L} = \begin{pmatrix} 0 & L_{1,2} & L_{1,3} & \dots & L_{1,T} \\ 0 & 0 & L_{2,3} & & L_{2,T} \\ | & & & & | \\ 0 & 0 & 0 & \dots & L_{T-1,T} \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

$$\mathcal{T} = \begin{pmatrix} B^{1}_{1} + S^{0}_{1} & 0 & \dots & 0 & 0 \\ 0 & B^{2}_{2} + S^{1}_{2} & \dots & 0 & 0 \\ | & & & & | \\ 0 & 0 & \dots & B^{T-1}_{T-1} + S^{T-2}_{T-1} & 0 \\ 0 & 0 & \dots & 0 & S^{T-1}_{T} \end{pmatrix}$$

Avec ces notations la matrice de variance-covariance se réécrit :

$$\mathcal{V}(\text{LInd}) = \sigma_G^2 \, (A'^{-1} \, \hat{I} \, A^{-1})^{-1} = \sigma_G^2 \, (\mathcal{T} - (\mathcal{L} + \mathcal{L}'))^{-1} \quad (25)$$

$\mathcal{V}$(LInd) s'exprime donc ici comme l'inverse d'une matrice facilement interprétable. L'annexe 18 établit que cette écriture peut être retransformée en :

$$\mathcal{V}(\text{LInd}) = \sigma_G^2 \, \mathcal{T}^{-1} \sum_{0}^{+\infty} [(\mathcal{L} + \mathcal{L}') \mathcal{T}^{-1}]^{i} \quad (26)$$

Les éléments des matrices $\mathcal{T}^{-1}$ et $\mathcal{L} + \mathcal{L}'$ étant tous positifs, il en est de même des composantes de $[(\mathcal{L} + \mathcal{L}') \mathcal{T}^{-1}]^{i}$, pour toutes les valeurs de i. Les variances et les covariances répertoriées par la matrice $\mathcal{V}$ (LInd) s'obtiennent alors comme des limites de séries de réels $\sum a_n$, où $a_n \geq 0$. On démontre donc, grâce à cette écriture, que les covariances sont toutes positives. La série des logarithmes des valeurs de l'indice estimé, $\ln(ind_t)$, aura alors tendance à être soit toujours au-dessus de la série des valeurs théoriques, $\ln(indice_t)$, soit toujours au-dessous. Les phénomènes d'oscillation autour des valeurs cibles seront donc un peu moins fréquents que les situations de surestimation ou de sous-estimation systématique, sans qu'il existe pour autant un biais pour cet estimateur.

2.6.4. La dynamique de l'indice

Le vecteur LInd = ($\ln(Ind_1)$, … , $\ln(Ind_T)$)' est un vecteur gaussien, centré sur les valeurs fondamentales ($\ln(Indice_1)$, … , $\ln(Indice_T)$)' et dont la matrice de variance-covariance est donnée par les formules (25) et (26).
On peut alors en déduire que chaque valeur Ind_t suit une loi log-normale. Le premier paramètre est $p_1 = \ln(Indice_t)$ et le second, noté p_2, s'obtient par une lecture directe de la diagonale de la matrice $\mathcal{V}$(LInd). Ind_t , estimateur de $Indice_t$, présente donc un biais positif[15] que l'on mesure par le coefficient multiplicatif $\exp(p_2 / 2)$. On retrouve ici les résultats empiriques du paragraphe 4.4, chapitre 2. Il pourrait être intéressant, dans de futures recherches, de rapprocher ces résultats théoriques des éléments fournis par Goetzmann et Peng dans leur article de 2002.

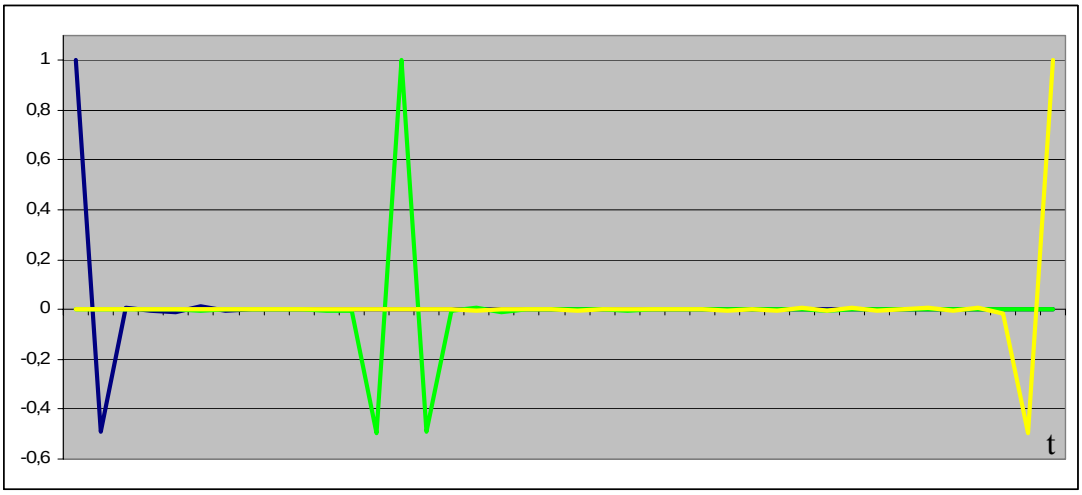
2.6.5. Un exemple numérique

Afin d'illustrer les formules présentées ci-dessus, on génère un échantillon de ventes répétées ω_0 en utilisant des valeurs classiques pour les paramètres[16]. On obtient une matrice d'information $\hat{I}$ dont on peut déduire la matrice des coefficients de corrélation du vecteur R, en normalisant $\mathcal{V}$(R) . La figure 1 présente les deux types de résultats que l'on rencontre alors. Pour les taux r_t associés à des intervalles non contigus aux bords ($t \neq 0$ et $t \neq 39$), le coefficient de corrélation de r_t avec lui-même vaut bien sûr 1, mais on remarquera surtout que les coefficients de r_t avec r_{t+1} et r_{t-1} sont très voisins[17] de - 0.5. Pour les autres dates ils sont par contre quasiment

[15] Si $X \sim \mathcal{LN}(p_1,p_2)$ alors $E(X) = \exp(p_1 + p_2/2)$
[16] La courbe des vrais prix est plate, la liquidité constante $K_i = 10000$, $\Theta = 10$ et $\sigma = 0{,}05$.
[17] De façon très intuitive la valeur de - 0,5 peut s'expliquer en remarquant qu'approximativement $r_t = \ln(p_{t+1}) - \ln(p_t)$ et $r_{t+1} = \ln(p_{t+2}) - \ln(p_{t+1})$. Si les prix sont générés indépendamment entre chaque date, la part commune entre r_t et r_{t+1}, c'est-à-dire $\ln(p_{t+1})$, se retrouve en d'égales proportions dans ces deux taux (au signe près). Le niveau de l'aléa commun doit donc être voisin de -0,5. Ce commentaire ne constitue en rien une démonstration, il s'agit simplement d'une tentative d'explication intuitive des observations, l'étude théorique pourrait être précisée sur ce point.

nuls. Pour r_0 et r_{39} la structure est presque identique, mais par construction on observera qu'une seule fois la valeur -0,5.

Figure 1 : Corrélogrammes des taux r_t



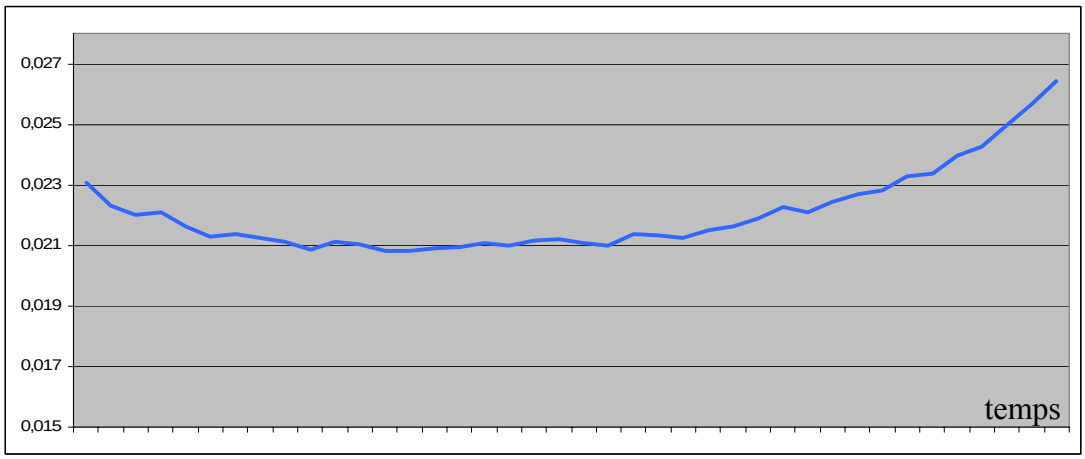
Séries des cœfficients de corrélation de r_k et r_t pour t = 0,…,39
Courbe bleue : k = 0 Courbe verte : k = 13 Courbe jaune : k = 39

Pour cet échantillon ω_0 il est également possible de déterminer l'ampleur théorique du biais évoqué dans le paragraphe 2.6.4. La figure 2 en donne la représentation graphique. On constate effectivement que l'indice de ventes répétées surestime les valeurs fondamentales, mais cette déviation reste très faible, en moyenne de l'ordre de 0,022%. Ces niveaux sont cohérents avec ceux observés dans le chapitre 2, paragraphe 4.4, où sous les mêmes conditions de simulation on détectait un biais empirique de 0,04%. L'ordre de grandeur est donc respecté. La différence entre ces deux niveaux est probablement due à un phénomène de fluctuation d'échantillonnage et à un nombre un peu limité d'itérations dans la simulation empirique.

La désagrégation des indicateurs empiriques du chapitre 2, paragraphe 4.4, n'avait pas permis de détecter une structure temporelle pour le biais. Avec la courbe théorique on constate par contre que la déviation n'est pas uniforme sur [0,T] : le biais est plus marqué aux bords de l'intervalle qu'au centre. De plus, alors que

l'erreur absolue exhibait une courbe en U symétrique[18] (chapitre 2, paragraphe 5.2.1), le bais se révèle être plus important à droite de l'intervalle qu'à gauche.

Figure 2 : Biais théorique des valeurs indicielles (en %) sur [0,40]



3. Réversibilité

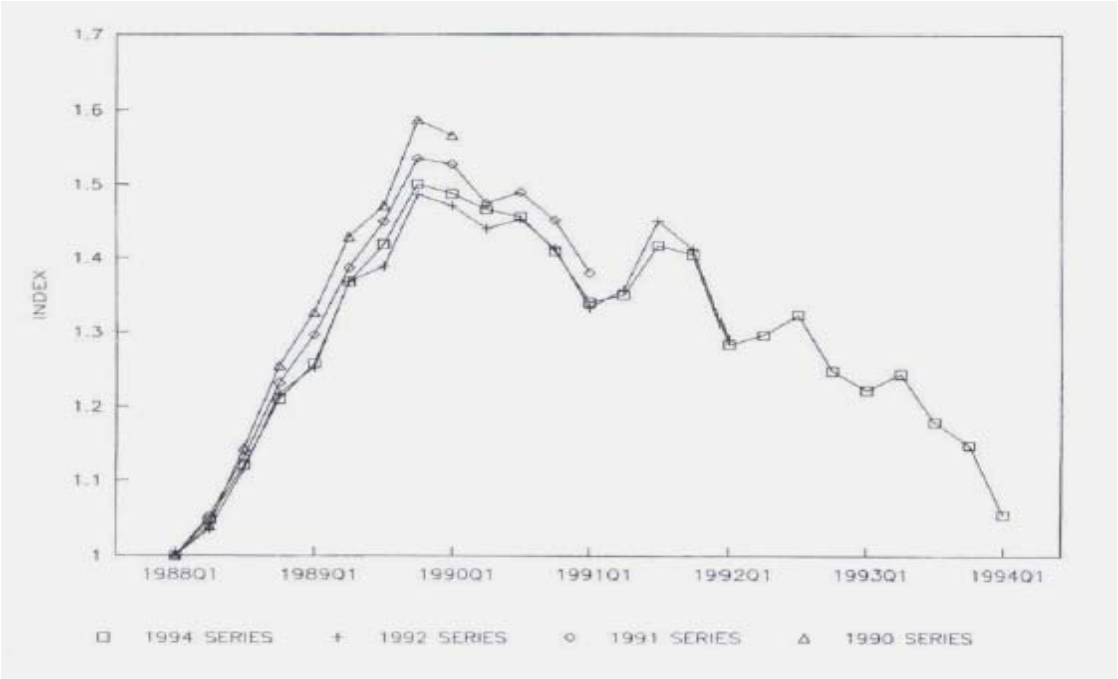
Ce paragraphe étudie et essaye de quantifier les conséquences d'un phénomène indésirable, propre au RSI : la réversibilité des valeurs passées. La démonstration de la formule théorique associée à ce phénomène se fera, à nouveau, en remontant des transactions aux valeurs indicielles par l'intermédiaire de la décomposition en briques élémentaires établie dans le chapitre 1. Certains passages, un peu techniques, pourront être survolés si l'on se concentre sur l'esprit de la démarche.

[18] Il est en fait probable que les erreurs présentent aussi une structure très légèrement asymétrique, car les erreurs intègrent le bais. Toutefois, comme l'échelle d'observation des erreurs est le pourcentage et que celle des biais est le centième de pourcentage, si l'asymétrie est une propriété découlant uniquement du biais, il n'est pas très surprenant de ne pas l'observer au niveau des erreurs absolues : cette caractéristique est trop minime pour pouvoir être détectée.

3.1. Le problème

Une des particularités du RSI est sa dépendance à l'horizon temporel d'estimation T. Une valeur Ind_t n'est en effet pas fixée définitivement, car si l'on étend l'horizon de T à T' (T'>T) de nouvelles ventes répétées pourront apporter des informations complémentaires[19] sur [0,T]. L'ancienne valeur estimée $Ind_t(T)$ ne sera pas nécessairement égale à la nouvelle, $Ind_t(T')$. Ce phénomène de volatilité rétroactive est illustré dans la figure 3, extraite de l'article de Clapp et Giaccotto (1999). Ces courbes indicielles concernent le comté de Los Angeles pour différents horizons (1990, 1991, 1992, 1994). Selon la fenêtre d'étude les valeurs produites pour la période commune [1988,1990] diffèrent.

Figure 3 : Comté de Los Angeles, Clapp, Giaccotto (1999)



[19] Par exemple, une donnée avec une date d'achat à t et une date de revente t', avec T < t' < T', sera informative pour [t,T]. Comme la revente s'est faite après T, elle n'aura cependant pas pu être prise en compte dans la première estimation de l'indice ; elle le sera par contre dans la deuxième.

La différence entre deux estimations pouvant être significative, il pourrait être intéressant d'essayer de quantifier les fluctuations potentielles, en utilisant le formalisme introduit précédemment.

3.2. Les notations

Dans toute cette section on supposera que l'horizon temporel, initialement T_1, est étendu à T_2 ($T_2 > T_1$), le tableau 2 illustre cette extension pour la distribution informationnelle. Les notations usuelles sont maintenues, mais l'horizon d'estimation sera ajouté en paramètre. Ainsi $H_p(t)$ sera noté $H_p(t;T_1)$ ou $H_p(t;T_2)$ selon les cas. Les grandeurs associées uniquement aux nouvelles ventes répétées apparues lors du passage de T_1 à T_2 seront, quant à elles, indexées par $T_2 \setminus T_1$; on écrira par exemple $H_p(t;T_2 \setminus T_1)$.

Tableau 2 : Nouvelles données associées à un changement d'horizon pour la distribution informationnelle

	0	1	…	t	t + 1	…	T_1	…	T_2
0		$L_{0,1}$	…	$L_{0,t}$	$L_{0,t+1}$	…	L_{0,T_1}	…	L_{0,T_2}
1			…	$L_{1,t}$	$L_{1,t+1}$	…	L_{1,T_1}	…	L_{1,T_2}
⋮				…	…	…	…	…	…
t					$L_{t,t+1}$	…	L_{t,T_1}	…	L_{t,T_2}
t + 1						…	L_{t+1,T_1}	…	L_{t+1,T_2}
⋮							…	…	…
T_1								…	L_{T_1,T_2}
⋮									…
T_2									

Traits pleins : date d'achat avant T_1 et revente après T_1 ($i < T_1 < j \leq T_2$)
Pointillés : date d'achat et date de revente entre T_1 et T_2 ($T_1 \leq i < j \leq T_2$)

La population des nouvelles ventes répétées peut être séparée en deux sous-populations : celles dont la date d'achat est avant T_1 et la revente après T_1 ($i < T_1 < j \leq T_2$), délimitées par les lignes continues dans le tableau 2, et celles dont la date d'achat et la date de revente sont toutes les deux entre T_1 et T_2 ($T_1 \leq i < j \leq T_2$), délimitées par les lignes en pointillées. Les ventes répétées pertinentes pour [t,t+1], si l'horizon est T_1, sont représentées en gris clair. Quand l'horizon devient T_2 il faut y rajouter les cellules en gris foncé.

On établira dans la suite de ce paragraphe les formules de réversibilité pour $H_p(t)$ et $H_f(t)$, puis pour ρ_t , pour $\hat{I}$, et enfin pour l'indice.

3.3. La réversibilité pour $H_p(t)$ et $H_f(t)$

3.3.1. La réversibilité pour I^t et n^t

Pour un intervalle de temps [t,t+1], avec $t < T_1$, les quantités d'information pertinentes sont :

$$I^t(T_1) = \sum_{i \leq t < j \leq T_1} L_{i,j} \qquad \text{pour le premier horizon}$$

$$I^t(T_2) = \sum_{i \leq t < j \leq T_2} L_{i,j} = I^t(T_1) + \sum_{i \leq t < T_1 < j \leq T_2} L_{i,j} \qquad \text{pour le deuxième horizon}$$

Pour le deuxième horizon, la somme indexée par $i \leq t < T_1 < j \leq T_2$ correspond à l'information complémentaire (en gris foncé dans le tableau 2) que l'on notera[20] $I^t(T_2 \setminus T_1)$. On obtient alors la relation :

$$I^t(T_2) = I^t(T_1) + I^t(T_2 \setminus T_1) \qquad (27)$$

[20] La notation $T_2 \setminus T_1$ peut être comprise comme "utile pour T_2 mais pas pour T_1"

Les équivalents réels de $I^t(T_2)$, $I^t(T_1)$, $I^t(T_2 \setminus T_1)$ seront notés $n^t(T_2)$, $n^t(T_1)$, $n^t(T_2 \setminus T_1)$ et on dispose alors d'une relation similaire :

$$n^t(T_2) = n^t(T_1) + n^t(T_2 \setminus T_1) \qquad (28)$$

3.3.2. La grandeur $H_p(t, T_2 \setminus T_1)$

Les quantités $[H_p(t)]^{I^t}$ se calculent avec les prix d'achats. Pour chacun des deux horizons, on a :

$$[H_p(t,T_1)]^{I^t(T_1)} = \prod_{i \le t < j \le T_1} (\Pi_{k'} p_{k',i})^{1/(\Theta + (j-i))} \quad \text{et} \quad [H_p(t,T_2)]^{I^t(T_2)} = \prod_{i \le t < j \le T_2} (\Pi_{k'} p_{k',i})^{1/(\Theta + (j-i))}$$

On peut alors écrire : $[H_p(t,T_2)]^{I^t(T_2)} = [H_p(t,T_1)]^{I^t(T_1)} \prod_{i \le t < T_1 < j \le T_2} (\Pi_{k'} p_{k',i})^{1/(\Theta + (j-i))}$

En utilisant les mêmes notations que dans le cas général (chapitre 1, paragraphe 6), ce produit devient :

$$\prod_{i \le t < T_1 < j \le T_2} (\Pi_{k'} p_{k',i})^{1/(\Theta + (j-i))} = \prod_{i \le t < T_1 < j \le T_2} (h_p^{(i,j)})^{L_{i,j}}$$

La masse totale de ces poids $L_{i,j}$ étant $I^t(T_2 \setminus T_1)$, on peut alors définir de manière naturelle la moyenne géométrique $H_p(t, T_2 \setminus T_1)$ par :

$$[H_p(t, T_2 \setminus T_1)]^{I^t(T_2 \setminus T_1)} = \prod_{i \le t < T_1 < j \le T_2} (h_p^{(i,j)})^{L_{i,j}} = \prod_{i \le t < T_1 < j \le T_2} (\Pi_{k'} p_{k',i})^{1/(\Theta + (j-i))} \qquad (29)$$

Pour l'intervalle [t, t+1], $H_p(t, T_2\backslash T_1)$ représente le prix moyen d'achat dans la population des nouvelles ventes répétées.

3.3.3. La formule de réversibilité pour $H_p(t)$ et $H_f(t)$

On peut maintenant écrire : $[H_p(t,T_2)]^{I^t(T_2)} = [H_p(t,T_1)]^{I^t(T_1)} [H_p(t, T_2\backslash T_1)]^{I^t(T_2\backslash T_1)}$ (30)

$H_p(t,T_2)$ n'est donc rien d'autre que la moyenne géométrique entre l'ancienne valeur $H_p(t,T_1)$ et le terme associé exclusivement aux nouvelles données $H_p(t, T_2 \setminus T_1)$, leur contribution respective étant mesurée par les poids informationnels $I^t(T_1)$ et $I^t(T_2\backslash T_1)$.

De même pour les prix de revente si l'on introduit les quantités :

$$[H_f(t, T_2 \setminus T_1)]^{I^t(T_2\backslash T_1)} = \prod_{i \le t < T_1 < j \le T_2} (h_f^{(i,j)})^{L_{i,j}} = \prod_{i \le t < T_1 < j \le T_2} (\Pi_{k'} p_{k',j})^{1/(\Theta + (j-i))} \quad (29')$$

On a: $[H_f(t,T_2)]^{I^t(T_2)} = [H_f(t,T_1)]^{I^t(T_1)} [H_f(t, T_2 \setminus T_1)]^{I^t(T_2 \setminus T_1)}$ (30')

3.4. La réversibilité pour la moyenne des taux moyens P

3.4.1. La réversibilité pour τ^t

Avant d'établir la formule pour la moyenne des taux moyens il est nécessaire de préciser le lien entre les périodes de détention moyennes $\tau^t(T_1)$ et $\tau^t(T_2)$. En

introduisant les grandeurs, $\zeta^t(T_2\backslash T_1)$ et $\tau^t(T_2\backslash T_1)$ pour les nouvelles ventes répétées, on démontre dans l'annexe 19 que :

$$\left[n^t(T_2)G(\zeta^t(T_2))\right]/\tau^t(T_2) = \left[n^t(T_1)G(\zeta^t(T_1))\right]/\tau^t(T_1) + \left[n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))\right]/\tau^t(T_2\backslash T_1) \quad (31)$$

La relation $n^t(T_2)G(\zeta^t(T_2)) = n^t(T_1)G(\zeta^t(T_1)) + n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))$ permet alors d'affirmer que $\tau^t(T_2)$ est simplement la moyenne harmonique pondérée de $\tau^t(T_1)$ et $\tau^t(T_2\backslash T_1)$.

3.4.2. La réversibilité pour ρ_t ($t < T_1$)

L'annexe 20 établit la formule de réversibilité[21] pour ρ_t quand $t < T_1$:

$$\rho_t(T_2) = [I^t(T_1)/I^t(T_2)][\tau^t(T_1)/\tau^t(T_2)]\,\rho_t(T_1) + [I^t(T_2\backslash T_1)/I^t(T_2)][\tau^t(T_2\backslash T_1)/\tau^t(T_2)]\rho_t(T_2\backslash T_1) \quad (32)$$

La quantité $I^t(T_1) / I^t(T_2)$ représente le pourcentage de l'information $I^t(T_2)$ déjà connue quand l'horizon est T_1 et $I^t(T_2\backslash T_1) / I^t(T_2)$ le pourcentage de l'information révélée entre T_1 et T_2. Les ratios $\tau^t(T_1) / \tau^t(T_2)$ et $\tau^t(T_2\backslash T_1) / \tau^t(T_2)$ mesurent les longueurs moyennes des périodes de détention pour les anciennes données et pour les nouvelles, relativement aux valeurs moyennes pour l'échantillon complété. Dans le cas particulier où les détentions moyennes sont égales pour $[0,T_1]$, $[0,T_2]$ et $[T_1, T_2]$, cette relation devient simplement :

$$\rho_t(T_2) = \left[I^t(T_1) / I^t(T_2)\right]\,\rho_t(T_1) + \left[I^t(T_2\backslash T_1) / I^t(T_2)\right]\,\rho_t(T_2\backslash T_1) \quad (32')$$

---

[21] On notera, pour $t < T_1$: $\rho_t(T_2\backslash T_1) = [\, n^t(T_2\backslash T_1)\, G(\zeta^t(T_2\backslash T_1))\,]^{-1} \sum_{i \le t < T_1 < j \le T_2} \sum_{k'} G(j - i)\, r_{k'}^{(i,j)}$

$= [(1/\tau^t(T_2\backslash T_1)) * (\ln H_f(t, T_2\backslash T_1) - \ln H_p(t, T_2\backslash T_1))]$

$\rho_t(T_2\backslash T_1)$ désigne la moyenne pondérée des taux moyens $r_{k'}^{(i,j)}$ pour les nouvelles ventes répétées

3.4.3. Généralisation des notations

Les formules (32) et (32') sont valables pour $t < T_1$ car, jusqu'à présent, les expressions définissant $I^t(T_2\backslash T_1)$, $\tau^t(T_2\backslash T_1)$, $\zeta^t(T_2\backslash T_1)$, $n^t(T_2 \setminus T_1)$ et $\rho_t(T_2\backslash T_1)$ n'ont été définies que pour $t < T_1$. On peut prolonger leur définition aux cas où $t \geq T_1$ de la façon suivante.

Pour ces grandeurs, les différentes sommes sont calculées originellement pour chaque cellule (i,j) telles que $i \leq t < T_1 < j \leq T_2$, c'est-à-dire pour toutes les nouvelles ventes répétées pertinentes pour [t,t+1] avec un $t < T_1$. Pour $t \geq T_1$, les cellules (i,j) pertinentes seront celles satisfaisant à la condition $i \leq t < j \leq T_2$.[22] Mais ce que l'on obtient alors n'est pas vraiment nouveau puisqu'il s'agit simplement de $I^t(T_2)$, $\tau^t(T_2)$, $\zeta^t(T_2)$, $n^t(T_2)$ et $\rho_t(T_2)$.

Ainsi par exemple, $I^t(T_2\backslash T_1) = \sum_{i \leq t < T_1 < j \leq T_2} L_{i,j}$ donnera pour $t \geq T_1$: $\sum_{i \leq t < j \leq T_2} L_{i,j}$, c'est-à-dire $I^t(T_2)$. Globalement, les diverses grandeurs de type $(T_2 \setminus T_1)$ pour $t \geq T_1$ sont donc simplement égales à celles obtenues pour l'échantillon T_2, pour ce même t ; à l'exception notable des $\{r_t\}$. Cette remarque va permettre d'écrire la formule (32) sous une forme plus synthétique.

3.4.4. L'écriture matricielle de la réversibilité pour P

Les taux $\rho_t(T_2)$, pour $0 \leq t < T_2$, sont regroupés dans le vecteur $P(T_2)$ de dimension T_2 ; les taux $\rho_t(T_1)$, pour $0 \leq t < T_1$, dans le vecteur $P(T_1)$ de dimension T_1. A partir du vecteur $P(T_1)$ on crée un vecteur de dimension T_2 en le complétant par $(T_2 - T_1)$ zéros, il sera noté en italiques par $P(T_1)$. Enfin les taux $\rho_t(T_2\backslash T_1)$

[22] $i \leq T_1 \leq t < j \leq T_2$ n'est pas correct car cela exclurait les ventes répétées pour lesquelles l'achat a été réalisé aux dates i telles que $T_1 < i \leq t$. Or, comme ces couples appartiennent aux nouvelles données et sont parfaitement pertinents pour [t,t+1] ils ne doivent pas être omis.

engendreront un vecteur de dimension T_2, noté $P(T_2\backslash T_1)$, dont les $(T_2 - T_1)$ dernières coordonnées seront simplement égales à celles de $P(T_2)$.

La relation (32), pour $t < T_1$, peut être écrite :

$$\tau^t(T_2)\, I^t(T_2)\, \rho_t(T_2) = I^t(T_1)\, \tau^t(T_1)\, \rho_t(T_1) + I^t(T_2\backslash T_1)\, \tau^t(T_2\backslash T_1)\, \rho_t(T_2\backslash T_1)$$

$$\Leftrightarrow n^t(T_2) G(\zeta^t(T_2))\, \rho_t(T_2) = n^t(T_1) G(\zeta^t(T_1))\, \rho_t(T_1) + n^t(T_2\backslash T_1) G(\zeta^t(T_2\backslash T_1))\, \rho_t(T_2\backslash T_1) \quad (33)$$

Et comme mentionné ci-dessus on a pour $t \geq T_1$, la formule triviale suivante :

$$n^t(T_2)\, G(\zeta^t(T_2))\, \rho_t(T_2) = n^t(T_2\backslash T_1)\, G(\zeta^t(T_2\backslash T_1))\, \rho_t(T_2\backslash T_1) \quad (33')$$

La matrice diagonale $\eta(T_1)$, dont les valeurs sont $n^0(T_1)G(\zeta^0(T_1)),\ldots, n^{T_1-1}(T_1)G(\zeta^{T_1-1}(T_1))$, peut être injectée dans une matrice de taille T_2 en la complétant pas des zéros ; on notera par $\eta(T_1)$ la matrice obtenue. $\eta(T_2)$ est une matrice diagonale de taille T_2 dont les valeurs sont $n^0(T_2)G(\zeta^0(T_2),\ldots, n^{T_2-1}(T_2)G(\zeta^{T_2-1}(T_2))$ et $\eta(T_2\backslash T_1)$ une matrice de même dimension construite avec les quantités $n^0(T_2\backslash T_1)G(\zeta^0(T_2\backslash T_1)$, … , $n^{T_2-1}(T_2\backslash T_1)G(\zeta^{T_2-1}(T_2\backslash T_1))$. Les deux types d'équations pour $t < T_1$ et $t \geq T_1$, formules (33) et (33'), peuvent maintenant s'écrire simultanément et simplement sous la forme :

$$\eta(T_2)\, P(T_2) = \eta(T_1)\, P(T_1) + \eta(T_2\backslash T_1)\, P(T_2\backslash T_1) \quad (33'')$$

3.5. La réversibilité pour Î

La formule de réversibilité pour Î ne nécessite pas de calculs compliqués. Pour un intervalle de temps $[t_i, t_j]$ l'information pertinente sera notée $I^{[t_i, t_j]}(T_1)$ ou $I^{[t_i, t_j]}(T_2)$ selon l'horizon considéré. Les matrices informationnelles associées, $\hat{I}(T_1)$ et

$\hat{I}(T_2)$, sont des matrices carrées de dimension T_1 et T_2 respectivement. Une troisième matrice carrée $\hat{I}(T_2\backslash T_1)$, de dimension T_2, va permettre de faire le lien entre $\hat{I}(T_1)$ et $\hat{I}(T_2)$. Ses valeurs se calculent à partir des nouveaux $L_{i,j}$ (cf. tableau 2) et elles représentent pour chaque intervalle $[t_i, t_j]$ le supplément d'information fournit par les nouvelles données. $\hat{I}(T_2\backslash T_1)$ peut se décomposer en trois sous-matrices $\hat{I}_a(T_2\backslash T_1)$, $\hat{I}_b(T_2\backslash T_1)$ et $\hat{I}_c(T_2\backslash T_1)$ de la façon suivante :

$$\hat{I}(T_2\backslash T_1) = \begin{pmatrix} \hat{I}_a(T_2\backslash T_1) & \hat{I}_b(T_2\backslash T_1) \\ {}^{t}\hat{I}_b(T_2\backslash T_1) & \hat{I}_c(T_2\backslash T_1) \end{pmatrix}$$

$\hat{I}_a(T_2\backslash T_1)$ et $\hat{I}_c(T_2\backslash T_1)$ sont des matrices carrées de dimension T_1 et $T_2 - T_1$, $\hat{I}_b(T_2\backslash T_1)$ est de taille $T_1*(T_2 - T_1)$ et sa transposée ${}^{t}\hat{I}_b(T_2\backslash T_1)$ de taille $(T_2 - T_1)*T_1$. De manière plus précise on a :

$$\hat{I}_a(T_2\backslash T_1) = \begin{pmatrix} I^{[0,T_1+1]}(T_2) & I^{[0,T_1+1]}(T_2) & I^{[0,T_1+1]}(T_2) & \cdots & I^{[0,T_1+1]}(T_2) \\ I^{[0,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & \cdots & I^{[1,T_1+1]}(T_2) \\ I^{[0,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & I^{[2,T_1+1]}(T_2) & \cdots & I^{[2,T_1+1]}(T_2) \\ \vdots & \vdots & \vdots & & \\ I^{[0,T_1+1]}(T_2) & I^{[1,T_1+1]}(T_2) & I^{[2,T_1+1]}(T_2) & & I^{[T_1-1,T_1+1]}(T_2) \end{pmatrix}$$

$$\hat{I}_b(T_2\backslash T_1) = \begin{pmatrix} I^{[0,T_1+1]}(T_2) & \cdots & I^{[0,T_2]}(T_2) \\ I^{[1,T_1+1]}(T_2) & \cdots & I^{[1,T_2]}(T_2) \\ I^{[2,T_1+1]}(T_2) & \cdots & I^{[2,T_2]}(T_2) \\ \vdots & & \vdots \\ I^{[T_1-1,T_1+1]}(T_2) & \cdots & I^{[T_1-1,T_2]}(T_2) \end{pmatrix}$$

$$\hat{I}_c(T_2\backslash T_1) = \begin{pmatrix} I^{[T_1,T_1+1]}(T_2) & \cdots & I^{[T_1,T_2]}(T_2) \\ \vdots & & \vdots \\ I^{[T_1,T_2]}(T_2) & \cdots & I^{[T_2-1,T_2]}(T_2) \end{pmatrix}$$

$\hat{I}_a(T_2\backslash T_1)$ est une matrice symétrique dont les éléments diagonaux correspondent à ceux de la première colonne de $\hat{I}_b(T_2\backslash T_1)$. A partir de ses éléments diagonaux, les valeurs de la matrice $\hat{I}_a(T_2\backslash T_1)$ sont les mêmes à droite et en dessous. En ce qui concerne les matrices $\hat{I}_b(T_2\backslash T_1)$ et $\hat{I}_c(T_2\backslash T_1)$ elles sont simplement extraites de la matrice $\hat{I}(T_2)$. $\hat{I}_a$ représente l'information additionnelle pour un intervalle $[t_i,t_j]$ inclut dans $[0,T_1]$: nouvelles données pour un intervalle ancien. $\hat{I}_c$ décrit l'information additionnelle pour un intervalle $[t_i,t_j]$ inclut dans $[T_1,T_2]$: nouvelles données pour un nouvel intervalle. Tandis que $\hat{I}_b$ concernent les intervalles $[t_i,t_j]$ inclut dans $[0,T_2]$

avec $T_1 \in]t_i,t_j[$: nouvelles données pour les intervalles qui enjambent le premier horizon T_1.

La matrice $\hat{I}(T_1)$ s'injecte naturellement dans une matrice carrée de dimension T_2, notée en italiques par $\hat{I}(T_1)$:

$$\hat{I}(T_1) = \begin{pmatrix} \hat{I}(T_1) & 0 \\ 0 & 0 \end{pmatrix}$$

La formule de réversibilité pour les matrices informationnelles s'écrit alors simplement :

$$\hat{I}(T_2) = \hat{I}(T_1) + \hat{I}(T_2 \backslash T_1) \qquad (34)$$

3.6. La réversibilité pour l'indice

Pour un horizon T_1, les concepts usuels $I^t(T_1)$, $\tau^t(T_1)$, $\zeta^t(T_1)$, $n^t(T_1)$ et $\rho_t(T_1)$ calculés pour $t < T_1$, engendrent un vecteur $R(T_1)$[23]. De même $I^t(T_2)$, $\tau^t(T_2)$, $\zeta^t(T_2)$, $n^t(T_2)$ et $\rho_t(T_2)$ calculés pour $t < T_2$, donnent un deuxième vecteur $R(T_2)$. Le lien entre ces deux vecteurs va être établi grâce aux quantités $I^t(T_2\backslash T_1)$, $\tau^t(T_2\backslash T_1)$, $\zeta^t(T_2\backslash T_1)$, $n^t(T_2\backslash T_1)$ et $\rho_t(T_2\backslash T_1)$. Si on examine précisément les définitions de ces grandeurs, nous pouvons remarquer qu'il ne s'agit en fait que des mesures I^t, τ^t, ζ^t, n^t et ρ_t que l'on obtient en restreignant les données d'estimation de l'indice de ventes

[23] Le vecteur de dimension T_1, $R(T_1)$, sera parfois complété avec $(T_2 - T_1)$ zéros pour obtenir un vecteur de dimension T_2, noté en italiques par $R(T_1)$

répétées sur $[0,T_2]$ aux nouvelles données, cf. tableau 3. Le calcul du RSI sur $[0,T_2]$, basé uniquement sur ce sous-échantillon, est alors immédiat car toutes les grandeurs intermédiaires sont déjà connues. On notera $R(T_2\backslash T_1)$ le vecteur obtenu[24].

Tableau 3 : Echantillon d'estimation pour $R(T_2\backslash T_1)$

	0	1	…	T_1	T_1+1	…	T_2
0		0	…	0	L_{0,T_1+1}	…	L_{0,T_2}
1			…	0	L_{1,T_1+1}	…	L_{1,T_2}
⋮				⋮	⋮	…	⋮
T_1					L_{T_1,T_1+1}	…	L_{T_1,T_2}
T_1+1						…	L_{T_1+1,T_2}
⋮							⋮
T_2							

En utilisant maintenant la relation $\hat{I}R = \eta P$ et l'équation (33''), la formule traduisant le phénomène de réversibilité pour l'indice s'écrit finalement sous une forme très simple :

$$\hat{I}(T_2)\, R(T_2) = \hat{I}(T_1)\, R(T_1) + \hat{I}(T_2\backslash T_1)\, R(T_2\backslash T_1) \qquad (35)$$

En résumé, la réversibilité indicielle s'exprime de la façon suivante :

- On estime dans un premier temps l'indice sur $[0,T_1]$ avec les données anciennes : on obtient une matrice d'information $\hat{I}(T_1)$ et un vecteur $R(T_1)$.
- Avec les nouvelles données, on estime le RSI sur $[0,T_2]$: on obtient $\hat{I}(T_2\backslash T_1)$ et $R(T_2\backslash T_1)$.
- Enfin, l'indice estimé sur $[0,T_2]$ en utilisant les nouvelles données et les anciennes donne $\hat{I}(T_2)$ et $R(T_2)$.

[24] On a pu voir ci-dessus que les grandeurs I^t, τ^t, ζ^t, n^t et ρ_t étaient égales pour T_2 et $T_2\backslash T_1$ si $t \geq T_1$. Malheureusement pour l'indice de ventes répétées ce type de relation n'est pas vérifiée : pour $t \geq T_1$ on n'aura pas en général $r_t(T_2\backslash T_1) = r_t(T_2)$.

Traduire la réversibilité revient alors simplement à affirmer que le produit ÎR est une grandeur additive lorsque l'horizon est étendu de T_1 à T_2. Cette formule (35) sera utilisée dans le paragraphe 3.8 pour quantifier concrètement l'ampleur potentielle du phénomène de réversibilité. Ses conséquences pour la pratique sont importantes.

3.7. Comparaison avec les résultats de Clapp, Giaccotto (1999)

Dans un contexte BMN, Clapp et Giaccotto ont formulé un résultat de réversibilité, susceptible d'être généralisé à une situation de type Case-Shiller. Comment situer la formule $\hat{I}(T_2)\ R(T_2) = \hat{I}(T_1)\ R(T_1) + \hat{I}(T_2 \backslash T_1)\ R(T_2 \backslash T_1)$ par rapport à celui-ci ? On utilisera les notations introduites dans le présent exposé et non pas celles de l'article original pour exposer leur résultat.

3.7.1. La formule de Clapp et Giacotto (1999)

Pour l'intervalle $[0,T_1]$ on effectue la régression $Y(T_1) = D(T_1)\ \text{LIndice}(T_1) + \varepsilon(T_1)$ où le vecteur à estimer n'est pas le vecteur des taux de croissance R^*, mais celui des valeurs de l'indice, LIndice[25]. Le vecteur $Y(T_1)$ répertorie les rendements logarithmiques réalisés par les ventes répétées de l'échantillon d'estimation. Les lignes de la matrice $D(T_1)$ sont associées à l'ensemble des transactions. Un +1 code la date de revente, un -1 la date d'achat et le reste de la ligne ne comporte que des 0 (le cas des achats à la date 0 est toujours traité de manière spécifique, comme dans le chapitre 1).

Quand on étend l'intervalle à $[0,T_2]$, la régression s'écrit $Y(T_2) = D(T_2)\ \text{LIndice}(T_2) + \varepsilon(T_2)$. On note alors :

[25] en logarithme

- $Y(T_2)' = (Y(T_1)' ; Y(T_2/T_1)')$: on ajoute aux anciennes données des rendements réalisés dans le premier échantillon, $Y(T_1)$, les nouveaux rendements $Y(T_2/T_1)$.

- $$D(T_2) = \begin{pmatrix} D(T_1) & 0 \\ D_1(T_2/T_1) & D_2(T_2/T_1) \end{pmatrix}$$

  La partie inférieure de la matrice bloc $D(T_2)$ concerne uniquement les nouvelles données. $D_1(T_2/T_1)$ répertorie les transactions effectuées avant T_1 (il ne s'agit ici que d'achats) et $D_2(T_2/T_1)$ celles réalisées après T_1 (achats et reventes). Les nouvelles données sont en fait de deux types : achat avant T_1 et revente après T_1, ou bien, achat et revente après T_1. Dans le premier cas, le -1 de l'achat est enregistré dans la matrice $D_1(T_2/T_1)$ et le +1 de revente dans $D_2(T_2/T_1)$; par contre dans le second cas, le -1 et le +1 sont inscrits ensemble dans $D_2(T_2/T_1)$. On notera par $\Delta(T_2) = (D(T_1)' ; D_1(T_2/T_1)')'$ la partie gauche de la matrice $D(T_2)$ et par $F(T_2) = (0' ; D_2(T_2/T_1)')'$ sa partie droite.

- Le vecteur des valeurs de l'indice $LIndice(T_2)$ couvre toutes les dates de $[0,T_2]$. On peut le décomposer en deux parties, la première donne les valeurs sur $[0,T_1]$, la seconde sur $]T_1 ,T_2]$:

$$LIndice(T_2)' = (LIndice_1(T_2)' ; LIndice_2(T_2)')$$

La formule de réversibilité de Clapp et Giaccotto établit le lien entre les vecteurs $LInd(T_1)$ et $LInd_1(T_2)$, qui fournissent tous les deux les valeurs d'un indice sur l'intervalle $[0,T_1]$. Le premier vecteur se sert uniquement de l'échantillon restreint, $Y(T_1)$, tandis que le second se sert de l'échantillon prolongé, $Y(T_2)$.

Cette formule fait intervenir une régression auxiliaire : $Y(T_2/T_1) = D_1(T_2/T_1)AUX+\varepsilon'$. Bien que cette régression ressemble à celles effectuées pour obtenir les différents indices, le vecteur AUX ne peut pas être interprété comme un indice ; il ne s'agit que du résultat abstrait d'un calcul. La formulation du résultat requiert, de plus, l'introduction de la matrice Ω définie par :

$$\Omega = \left[D(T_1)' D(T_1) + D_1(T_2/T_1)' D_1(T_2/T_1) \right]^{-1} D(T_1)' D(T_1) \qquad (36)$$

La formule de réversibilité s'écrit alors :

$$LInd_1(T_2) = \Omega\, LInd(T_1) + (I-\Omega)\, AUX + \left[\Delta(T_2)'\Delta(T_2)\right]^{-1}\Delta(T_2)'F(T_2)LInd_2(T_2) \qquad (37)$$

3.7.2. Comparaison des formules de réversibilité

Si l'on choisit de travailler avec les concepts introduits dans ce travail, la formule de réversibilité apparaît sous une forme très simple, que l'on rappelle ci-dessous :

$$\hat{I}(T_2)\, R(T_2) = \hat{I}(T_1)\, R(T_1) + \hat{I}(T_2\backslash T_1)\, R(T_2\backslash T_1)$$

Ces deux formules sont bien sûr équivalentes théoriquement, car elles traduisent le même phénomène, mais sous l'angle de la pratique elles sont très différentes. Celle de Clapp et Giaccotto est assez lourde à écrire ou à manipuler et difficilement interprétable en termes financiers. Que représente par exemple la matrice Ω ? La régression auxiliaire $Y(T_2/T_1) = D_1(T_2/T_1)\, AUX + \varepsilon'$ est de plus un calcul abstrait qui ne correspond pas à un indice intermédiaire.

La formule concurrente est plus simple. Elle ne fait intervenir que les matrices d'information et les taux de croissance des indices, ces différentes

grandeurs sont toutes interprétables. L'équivalent de la régression auxiliaire AUX n'est plus ici un artifice de calcul, car elle correspond à l'estimation de l'indice $R(T_2\backslash T_1)$ sur l'intervalle $[0,T_2]$, réalisé en utilisant uniquement les nouvelles données. Il semble donc que la formule établie dans ce paragraphe soit plus performante que celle de Clapp et Giaccotto.

On pourrait d'ailleurs la comprendre comme une sorte « d'équation de conservation de l'énergie » du système, où cette notion d'énergie serait définie par le produit IR. La formule de réversibilité s'exprimerait alors intuitivement par :

Energie fournie par les données étendues $\hat{I}(T_2)\,R(T_2)$	=	**Energie fournie par les données restreintes** $\hat{I}(T_1)\,R(T_1)$	+	**Energie fournie par les nouvelles données** $\hat{I}(T_2\backslash T_1)\,R(T_2\backslash T_1)$

Ce concept d'énergie fournie par un lot de données pourrait aussi servir à interpréter la relation $\hat{I}R = \eta P$. Le terme de gauche pourrait être vu comme la quantité d'énergie du système des valeurs de l'indice (système informationnel) et le terme de droite comme la quantité d'énergie fournie par les données d'estimation (système réel). Là encore il s'agit d'une équation de conservation de l'énergie que l'on peut résumer par :

Energie du système informationnel $\hat{I}\,R$	=	**Energie fournie par le système réel** $\eta\,P$

On pourra trouver dans l'annexe 23 une illustration numérique de la formule de réversibilité.

3.8. Une quantification empirique de la réversibilité

Le problème de la réversibilité de l'indice a été résolu sur le plan théorique par l'obtention de la formule (35). La volatilité rétroactive du RSI provient du fait que l'on ne travaille pas directement sur le passé mais sur la connaissance que l'on en a, et que cette compréhension s'améliore au fur et à mesure que de nouvelles données deviennent observables. Ce mécanisme de modification des croyances concernant le passé peut être un obstacle à l'introduction de produits dérivés indiciels et, plus généralement, il constitue une caractéristique indésirable pour la gestion du risque immobilier. Il est donc souhaitable, à partir de cette formule théorique, de pouvoir mettre en œuvre une méthodologie empirique permettant d'anticiper l'ampleur des révisions éventuelles. Il n'existe pas actuellement dans la littérature de technique comparable quantifiant explicitement ce phénomène.

On illustrera dans ce paragraphe, sur un cas concret, le problème posé par la réversibilité. Le principe de simulation sera ensuite présenté et commenté. Enfin, on appliquera cette méthode au cas exposé initialement pour pouvoir comparer le résultat réel et le résultat obtenu par simulation.

3.8.1. Un exemple

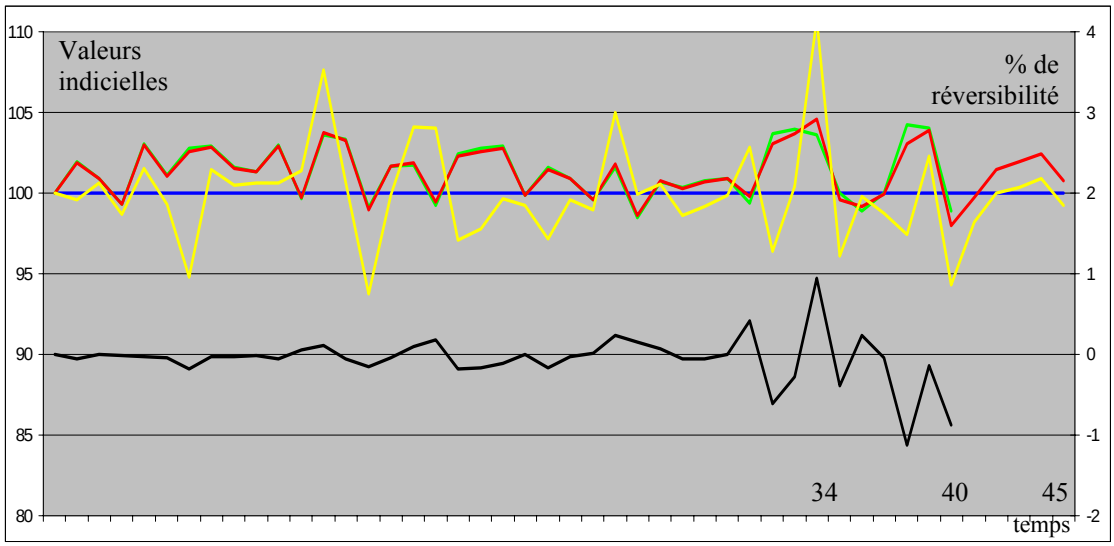
Reprenons le générateur de données, et sous des conditions standard[26], faisons passer l'horizon d'estimation de T = 40 à T = 45. A partir d'un échantillon prolongé, on calcule l'indice sur [0,45]. Puis, en le tronquant, nous calculons l'indice sur [0,40] avec les seules données observables à la date t = 40. Un troisième indice est calculé sur [0,45], en ne se servant que des nouvelles données[27]. La figure 4 est un exemple représentatif des résultats que l'on obtient dans cette situation. Les différentes estimations correspondent respectivement aux courbes rouge, verte et

[26] $\Theta = 10$; $\sigma = 0{,}05$; $K_i = 1$ et $\mathcal{K}_0 = 10000$; courbe des vrais prix plate au niveau 100

[27] Les nouvelles données sont celles qui se rajoutent à l'échantillon d'estimation lorsque l'on passe de T = 40 à T = 45.

jaune (les "vrais prix" correspondent à la courbe bleue). La courbe noire, associée à l'axe de droite, donne le pourcentage de variation entre l'ancien indice (vert) et l'indice complété (rouge), en prenant les anciennes valeurs de [0,40] comme référence[28].

Figure 4 : Un exemple de réversibilité



Axe de gauche : courbes indicielles — Axe de droite : réversibilité en %
Courbe bleue : vrais prix — Courbe verte : T = 40 — Courbe rouge : T = 45
Courbe jaune : indice nouvelles données et verte — Courbe noire : réversibilité en % entre les courbes rouge

L'échantillon des nouvelles données étant moins volumineux que ceux associés aux courbes verte et rouge, la courbe jaune est logiquement plus volatile. Pour la majorité des dates il est difficile de détecter une différence entre les valeurs de l'indice ancien et celles de l'indice complété (la courbe noire est très proche de 0). Ce n'est que vers le dernier quart de l'intervalle [0,40] que l'on commence à voir une certaine divergence[29] pouvant aller jusqu'à 1%. Le sens de cette divergence est déterminé par les nouvelles données. Par exemple à la date t = 34, celles-ci produisent des prix immobiliers élevés (110), mais les anciennes sont par contre à un

[28] $100\,(\,Ind_t(45) / Ind_t(40) - 1\,)$ pour $t = 0,\ldots,40$

[29] Cette faible divergence est due à la taille importante des échantillons ($\mathcal{K}_0 = 10000$). En pratique le phénomène de réversibilité pourra être plus important.

niveau plus faible (104). La courbe jaune amène alors la courbe rouge à un niveau supérieur à celui de l'ancienne estimation (voisin de 105).

La réversibilité présente donc une structure temporelle marquée. Elle semble se manifester, pour l'essentiel, aux dates les plus récentes. Or les valeurs de ce passé proche sont bien souvent les plus importantes dans une perspective d'investissement. Il est donc primordial d'élaborer une méthodologie de quantification, afin de déterminer le niveau de fiabilité des dernières valeurs indicielles (en d'autres termes, on cherche un intervalle de confiance).

3.8.2. Le principe de simulation

La quantification empirique de la réversibilité peut être réalisée par l'intermédiaire de simulations de type Monte-Carlo. L'algorithme de calcul est présenté dans la figure 5. A partir d'un échantillon de données réelles sur $[0,T_1]$ on calcule l'indice correspondant, ou plus exactement son vecteur des taux de croissance $R(T_1)$, ainsi que la matrice informationnelle associée[30] $\hat{I}(T_1)$. Ces deux grandeurs seront fixes pendant toute la simulation. Le but poursuivi ici consistera à déterminer la réversibilité sur $R(T_1)$, ou plus exactement sur $Ind(T_1)$, lorsque l'horizon est étendu de T_1 à T_2.

En prenant pour référence la distribution de l'information sur $[0,T_1]$, la première étape de l'algorithme consiste à étalonner le benchmark exponentiel[31] en déterminant les niveaux de K (le flux constant sur le marché) et α (la vitesse de revente). Lorsque l'horizon d'estimation passe de T_1 à T_2, on suppose que la

[30] Pour mémoire le vecteur $R(T_1)$, de dimension T_1, est complété par des zéros pour obtenir le vecteur $R(T_1)$ de dimension T_2 . De même la T_1-matrice $\hat{I}(T_1)$, complétée par des zéros donne la T_2-matrice $\hat{I}(T_1)$

[31] Par exemple sur le couple (N , I) = (nombre total de données, quantité totale d'information) dans l'échantillon.

distribution des nouvelles ventes répétées[32] est assimilable à celle du benchmark étalonné. Nous pouvons alors en déduire la matrice $\hat{I}(T_2\backslash T_1)$, puis la matrice $\hat{I}(T_2)$ en sommant $\hat{I}(T_1)$ et $\hat{I}(T_2\backslash T_1)$.

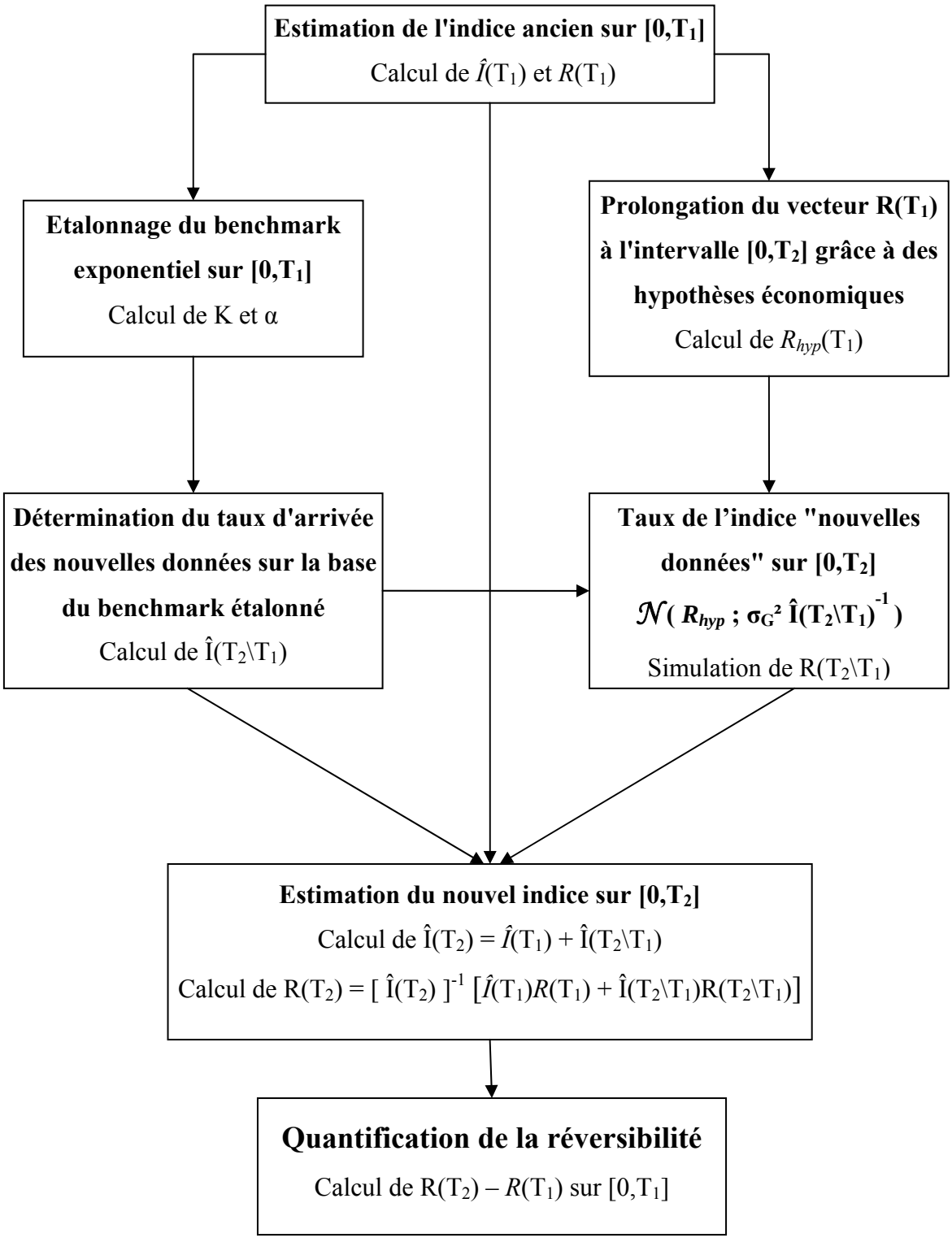
Le vecteur des taux $R(T_1)$ ne renseigne que sur l'intervalle $[0,T_1]$. En posant des hypothèses économiques sur l'évolution des prix immobiliers sur $[T_1,T_2]$, on le complète en un vecteur[33] R_{hyp} = ($R(T_1)$, $R_{hyp}(T_1;T_2)$). D'après le chapitre 3 paragraphe 7.5, le vecteur $R(T_2\backslash T_1)$ est un vecteur gaussien centré sur les taux de croissance monopériodiques de l'indice théorique. Sa matrice de variance-covariance est $\sigma_G^2\ \hat{I}(T_2\backslash T_1)^{-1}$. A la date T_1 son observation étant impossible, on le génèrera aléatoirement comme un vecteur gaussien $\mathcal{N}\left(R_{hyp} ; \sigma_G^2\ \hat{I}(T_2\backslash T_1)^{-1} \right)$. L'espérance théorique est remplacée ici par le meilleur estimateur dont on dispose à la date T_1 pour cette grandeur, à savoir $R(T_1)$ prolongé sur $[T_1,T_2]$ par des hypothèses économiques idoines.

Nous pouvons maintenant obtenir le vecteur $R(T_2)$, qui est l'unique inconnue dans la relation théorique de réversibilité $\hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1) = \hat{I}(T_2)R(T_2)$, en résolvant cette équation. Pour chacune de ces simulations l'amplitude de la modification se mesurera sur les T_1 premières composantes de la différence $R(T_2) - R(T_1)$ ou, de manière équivalente, sur les valeurs cumulées des indices.

[32] Celles dont la revente se produit après T_1 et au plus tard à T_2

[33] R_{hyp} peut aussi s'écrire $R(T_1) + R_{hyp}(T_1;T_2)$ où $R_{hyp}(T_1;T_2)$ = (0, $R_{hyp}(T_1;T_2)$). Le 0 est un vecteur de dimension T_1 ne comportant que des zéros et $R_{hyp}(T_1;T_2)$ un vecteur de dimension $T_2 - T_1$ associé aux hypothèses économiques sur l'intervalle $[T_1,T_2]$

Figure 5 : Algorithme de quantification de la réversibilité (de $[0,T_1]$ à $[0,T_2]$)



3.8.3. Commentaires

Dans le processus, tel qu'il a été présenté ci-dessus, l'aléa intervient à un seul endroit : lors de la génération du vecteur gaussien $R(T_2 \backslash T_1)$. On utilisera pour réaliser cette simulation un résultat classique d'algèbre linéaire, la décomposition de Cholesky.

Théorème (Girardin, Limnios p 134)

1. Si Γ est une matrice carrée de dimension d, symétrique, positive, de rang r
   Alors il existe une matrice B de dimension d x r et de rang r telle que : $\Gamma = B\,B'$ (Décomposition de Cholesky)

2. Soit : M un vecteur de dimension d
   Γ une matrice carrée de dimension d, symétrique, positive, de rang r
   $\Gamma = B\,B'$ la décomposition de Cholesky de Γ
   Si $Y \sim \mathcal{N}(0 , I_d)$ Alors $M + B\,Y \sim \mathcal{N}(M , \Gamma)$

Si on le souhaite, il est possible d'approfondir un peu plus la simulation en introduisant deux autres sources d'aléas. La première concernerait les $(T_2 - T_1)$ dernières coordonnées du vecteur R_{hyp} et la seconde le couple (K,α).

Pour estimer l'espérance du vecteur $R(T_2 \backslash T_1)$, le vecteur $R(T_1)$ a été prolongé en le vecteur R_{hyp}, en anticipant un scénario économique pour l'intervalle $[T_1,T_2]$. Toutefois, comme l'avenir est incertain, il pourrait être plus raisonnable d'autoriser les dernières coordonnées de R_{hyp} à varier aléatoirement plutôt que de se restreindre à une seule trajectoire prévisionnelle.

La deuxième généralisation envisageable porterait sur le couple (K,α). En étalonnant ces valeurs sur le taux d'arrivée de l'information observé au cours de $[0,T_1]$, nous avons déterminé des paramètres (K_0,α_0). Or il est tout à fait possible que sur

l'intervalle $[T_2,T_1]$ le rythme des transactions s'écarte de la moyenne. Pour traduire cette éventualité on pourrait alors autoriser K à varier aléatoirement dans un intervalle du type $[K_0 - \varepsilon ; K_0 + \varepsilon]$ et α_0 dans $[\alpha_0 - \varepsilon' ; \alpha_0 - \varepsilon']$.

L'analyse pourrait même être poursuivie plus avant en considérant que le rythme des transactions dépend de l'environnement économique, et donc des valeurs futures des prix immobiliers. Il s'agirait alors d'étalonner un modèle de hasard proportionnel[34] sur $[0,T_1]$, comme celui présenté par Cheung, Yau, Hui (2004), et en fonction du scénario simulé sur $[T_1,T_2]$ d'en déduire le rythme des ventes répétées.

3.8.4. Un exemple (suite)

Sous les conditions présentées dans le 3.8.1, on génère un échantillon ω_0 de ventes répétées pour l'horizon T_1 et on calcule l'indice associé. Puis, en utilisant le benchmark exponentiel étalonné sur $[0,T_1]$, on simule la matrice $\hat{I}(T_2\backslash T_1)$. Les quantités $\hat{I}(T_1)$, $R(T_1)$, $\hat{I}(T_2\backslash T_1)$, $\hat{I}(T_2)$ sont donc fixes et le vecteur $R(T_2\backslash T_1)$ sera la seule source d'aléa. On appliquera ici strictement la méthodologie du 3.8.2 en supposant que $R_{hyp}(T_1 ;T_2)$ n'est pas aléatoire.

Dans ce contexte simplifié l'étude formelle de la réversibilité peut être poussée plus avant. En utilisant la formule $R(T_2) = [\hat{I}(T_2)]^{-1}[\hat{I}(T_1)R(T_1)+\hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)]$ on démontre (cf. annexe 21) que le vecteur $R(T_2)$ est un vecteur gaussien pour lequel:

$$E[R(T_2)] = R(T_1) + \left[[\, \hat{I}(T_2)]^{-1} \hat{I}(T_2\backslash T_1)\right] R_{hyp}(T_1;T_2) \qquad (38)$$

$$\mathcal{V}[R(T_2)] = \sigma_G^2 \left[[\, \hat{I}(T_2)]^{-1} \hat{I}(T_2\backslash T_1)\right] [\, \hat{I}(T_2)]^{-1} \qquad (39)$$

La matrice $\hat{I}(T_2\backslash T_1)$ représente l'information nouvelle et la matrice $\hat{I}(T_2)$ l'information totale. Le produit matriciel $\left[[\, \hat{I}(T_2)]^{-1} \hat{I}(T_2\backslash T_1)\right]$ qui apparaît dans ces

[34] Modèle de Cox. On pourra consulter Simon (2004) pour une présentation de cette technique économétrique.

deux formules peut donc s'interpréter comme la proportion (vectorielle) d'information nouvelle dans l'information totale.

La première relation indique que l'espérance de $R(T_2)$ est égale à l'ancien vecteur $R(T_1)$ plus une quantité traduisant l'influence des hypothèses économiques faites sur l'intervalle $[T_1,T_2]$, retranscrites par l'intermédiaires du vecteur $R_{hyp}(T_1;T_2)$. Cette influence est pondérée par le produit matriciel $\left[\left[\hat{I}(T_2)\right]^{-1} \hat{I}(T_2 \backslash T_1)\right]$ qui mesure le poids informationnel relatif des nouvelles données.
La deuxième formule doit, quant à elle, être rapprochée de la formule générale $\mathcal{V}[R(T_2)] = \sigma_G^2 \left[\hat{I}(T_2)\right]^{-1}$ qu'il faudrait appliquer si l'on souhaitait calculer directement l'indice sur $[0,T_2]$, sans chercher à faire une première estimation sur $[0,T_1]$. Lorsque l'on se place dans une situation de réversibilité, on suppose qu'une partie des transactions sont connues ; l'indice qui en résulte est donc logiquement moins volatil. La formule (39) indique en fait que ce coefficient d'atténuation de la volatilité n'est rien d'autre que $\left[\left[\hat{I}(T_2)\right]^{-1} \hat{I}(T_2 \backslash T_1)\right]$, à nouveau.

Si l'on décide maintenant de considérer, vu de la date T_1, que sur l'intervalle $[T_1,T_2]$ les prix de l'immobilier resteront constants. Ou, en d'autres termes, si l'on suppose que le vecteur $R_{hyp}(T_1;T_2)$ est égal au vecteur nul, on peut démontrer (cf. annexe 21) le résultat suivant :

Loi de réversibilité théorique

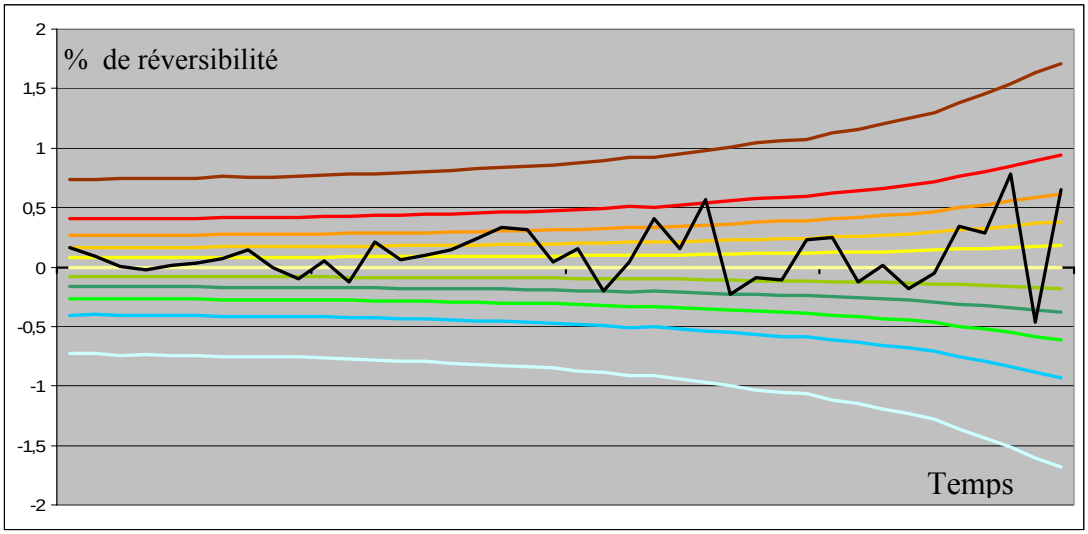
Pour $t = 1,\ldots,T_1$ le rapport $Ind_t(T_2) / Ind_t(T_1)$ suit une loi log-normale $\mathcal{LN}(0; v(t))$ où $v(t)$ est le t^{eme} élément diagonal de la matrice[35] :

$$\sigma_G^2 A(T_2)\left[\left[\hat{I}(T_2)\right]^{-1} \hat{I}(T_2 \backslash T_1)\right]\left[\hat{I}(T_2)\right]^{-1} [A(T_2)]'$$

[35] La matrice $A(T_2)$ est de dimension T_2, elle est constituée de 1 sur sa diagonale et en dessous, et de 0 ailleurs.

Le pourcentage de réversibilité[36] pour la date t est donc une variable aléatoire qui s'écrit $100*(Y_t - 1)$, où $Y_t \sim \mathcal{LM}(0; v(t))$. La figure 6 représente les différents déciles théoriques que l'on obtient pour ces pourcentages, à partir d'un échantillon ω_0. Dans ce graphique la courbe noire mesure simultanément la réversibilité empirique[37] observée lors de la simulation de cet échantillon ω_0. L'ampleur des révisions potentielles est faible et approximativement constante pour les deux premiers tiers de l'intervalle. Lorsque l'on se rapproche de l'horizon de la première estimation (T = 40), les fluctuations peuvent devenir plus importantes, comme en témoigne la divergence des courbes théoriques de la figure 6.

Figure 6 : Déciles des pourcentages de réversibilité (t = 1, …,40)



La courbe noire donne la réversibilité empirique pour ω_0. Les deux courbes extrêmes représentent les centiles à 1% et 99% Les neuf courbes intérieures représentent les déciles, de 10% à 90%

En appliquant la méthodologie présentée dans ce paragraphe, il est donc maintenant possible d'anticiper et de quantifier les effets de la réversibilité, d'une manière relativement fiable. Les courbes théoriques rendent compte, en effet, très correctement des courbes empiriques.

[36] $100*(Ind_t(T_2) / Ind_t(T_1) - 1)$

[37] Un indice est d'abord calculé sur [0,40] en utilisant uniquement les données observables dans ω_0 pour cette période, puis l'ensemble des ventes répétées est utilisé pour calculer un indice sur [0,45]. La réversibilité empirique correspond à la variation des valeurs indicielles sur [0,40] entre ces deux estimations.

4. Le problème des deux populations

Cette section examine les conséquences de l'estimation d'un indice unique lorsque l'échantillon des ventes répétées est constitué de deux sous-groupes ayant des dynamiques distinctes. On précisera dans un premier temps les hypothèses économiques implicites au modèle du benchmark exponentiel, en termes de flux et de stock. Le « problème des deux populations » sera ensuite exposé. En utilisant les propriétés du benchmark, on mènera tout d'abord une analyse théorique, dans laquelle la notion centrale sera l'effet de bord. Puis une étude plus empirique sera réalisée. Cette dernière partie mettra en lumière les concepts de population dominante et de population dominée.

4.1. Flux et stock pour le benchmark exponentiel

Dans ce paragraphe nous vérifierons l'égalité offre/demande pour le benchmark, nous déterminerons le flux et le stock associés à cette modélisation et nous étudierons leur rapport. Ces résultats seront réemployés par la suite.

4.1.1. L'égalité offre / demande

Considérons une population de biens parfaitement homogène dont le comportement correspond exactement à celui du benchmark exponentiel :

- K biens sont échangés sur le marché à chaque date t (flux constant)
- La loi de revente suit une distribution exponentielle de paramètre λ, λ est indépendant du temps

Dans le calcul d'un indice de ventes répétées, les transactions considérées sont celles dont la période de détention est incluse dans [0,T]. Celles dont l'achat est avant 0 ou

la revente après T ne sont pas prises en compte. Si l'on veut étudier, le marché immobilier dans toute sa généralité temporelle, il faut introduire une modélisation sans barrières de temps, c'est-à-dire examiner le comportement de toutes les cohortes[38] entre -∞ et + ∞ . Ceci revient alors à considérer, de manière fictive, que le marché immobilier existe de toute éternité. Mais, avant de pouvoir raisonner à ce niveau, il faut au préalable s'assurer que le modèle est cohérent. A une date t, K biens sont achetés (demande). La loi de revente des cohortes antérieures doit donc permettre d'assurer K reventes à cette même date (offre). S'il n'y a pas égalité entre l'offre et la demande le modèle est contradictoire.

Pour vérifier ce point on rappelle que pour les actifs achetés à t' (- ∞ < t' < t) le nombre de reventes à t vaut $K'\alpha^{t-t'}$. L'offre totale[39] de biens à t est alors :

$$Flux_t = \sum_{-\infty < t' < t} K'\alpha^{t-t'} = K' \sum_{t'' = 1, \dots, +\infty} \alpha^{t''} = K' \, \alpha / (1 - \alpha) = K \qquad (40)$$

L'offre à t correspond à la demande à t, le modèle est donc bien cohérent.

4.1.2. Stock et hypothèses économiques implicites

Dans le même ordre d'idée, le parc immobilier à la date t (stock) peut se calculer[40] comme la somme des biens achetés à t' (t' < t) et non encore revendus à t, auquel il faut ajouter les reventes réalisées à t.

$$Stock_t = \sum_{-\infty < t' < t} K \alpha^{t-t'} + K = K \, \alpha / (1 - \alpha) + K = K / (1 - \alpha) \qquad (41)$$

[38] ensemble de biens négociés sur le marché à une date donnée
[39] $K' = K(1 - \alpha) / \alpha$ et $\alpha = e^{-\lambda}$.
[40] Ce calcul signifie simplement que tous les biens du parc ont un jour été achetés

Le stock ne dépend donc pas du temps. Ce résultat très simple nous amène à préciser certaines hypothèses économiques implicites au modèle du benchmark exponentiel. Un stock constant signifie qu'il n'y a pas de nouvelles constructions ni de destructions de biens (pour raison d'obsolescence par exemple). D'un point de vue pratique, il serait plus exact de considérer que les nouveaux biens compensent exactement le nombre des démolitions, mais comme cette question n'est pas prise en compte dans la modélisation du benchmark, cela signifie indirectement que l'on travaille sur un parc immobilier éternel. Cette remarque ouvre une possibilité intéressante d'approfondissement pour le modèle. On pourrait, en effet, imaginer que $Stock_t$ doive satisfaire à l'équation plus générale:

$$Stock_t = Stock_{t-1} + (\text{nouvelles constructions})_t - (\text{destructions})_t \qquad (42)$$

et intégrer la variabilité du parc dans la dynamique des flux.

4.1.3. Le rapport flux / stock

D'après ce qui précède le taux de rotation des biens, $Flux_t / Stock_t$, est constant et vaut $1 - \alpha$. Comme $1 - \alpha = 1 - e^{-\lambda} \approx \lambda$, on a donc $Flux_t \approx \lambda\ Stock_t$. Ce constat s'accorde avec la spécification exponentielle de la loi de revente puisque le taux de revente instantané[41] vaut aussi λ, mais il ne constitue pas pour autant un résultat complètement trivial. La loi exponentielle s'applique en effet aux cohortes (effectifs négociés sur le marché à chaque date) dont elle modélise le devenir, tandis que le résultat sur le flux et le stock concerne l'ensemble des biens existants à un instant donné, indépendamment de leur date d'origine. Dans le cas du benchmark exponentiel il est donc établi que le comportement probabiliste est le même pour ces deux niveaux. On pourrait cependant tout à fait imaginer un autre type de benchmark fonctionnant différemment.

[41] Pourcentage de reventes réalisées sur le prochain intervalle de temps unitaire, en proportion de l'effectif toujours en vie.

4.2. La sur-représentativité de certains actifs immobiliers dans l'échantillon d'estimation

Si l'on cherche à estimer un indice de ventes répétées sur l'intervalle [0,T], le principe de sélection des données doubles (achat et revente) peut amener certaines complications. Supposons par exemple qu'il existe deux types de biens, les premiers sont détenus pendant une unité temporelle et les seconds pendant 10 unités de temps (la détention est parfaitement déterministe ici). Supposons de plus qu'à chaque date le nombre de transactions effectuées dans chacune des sous-populations est constant et fixé à 100.

Pour l'intervalle [0,1], les ventes répétées pertinentes seront de deux types : 100 biens à détention courte (achat à t_0, revente à t_1) et 100 biens à détention longue (achat à t_0, revente à t_{10}). Pour l'intervalle [1,2] il y aura toujours 100 biens à détention courte (achat à t_1, revente à t_2) mais le nombre de biens à détention longue sera cette fois de 200 (achat à t_0 et revente à t_{10}, ou achat à t_1 et revente à t_{11}). L'importance relative des effectifs pertinents de ces deux types de biens passera donc d'un facteur 1 à un rapport de 0.5, puis 0.33 , 0.25 (…) pour les intervalles suivants. Ce calcul n'a pas pour but de décrire précisément le phénomène, car il faudrait alors parler de détention probablement courte et de détention probablement longue. On cherche simplement ici à mettre en évidence la sur-représentativité mécanique de la famille des biens rapidement revendus dans l'échantillon d'estimation.

Ce problème a déjà été mentionné dans la littérature, on pourra par exemple consulter Clapp, Giaccotto (1999) sur ce sujet. Cet article traite du problème plus global de la réversibilité mais le biais de sélection induit par la méthode des ventes répétées est présenté de manière claire. Les auteurs emploient le terme de « flip » pour les détentions courtes. Si les deux familles de biens (détention longue, détention courte) suivaient la même dynamique de prix, cette particularité pourrait être négligée. Malheureusement, dans la pratique les « flips » présentent souvent des taux

de rentabilité supérieurs. En conséquence leur surnombre près du bord gauche de l'intervalle [0,T], a tendance à produire une surévaluation mécanique de l'indice pour les premières dates.

4.3. Les effets de bords dans un modèle à deux populations

Comme on a déjà pu le pressentir, la surreprésentation des « flips » diffère entre les bords de l'intervalle et son centre. On mènera ici une analyse plus approfondie de ce phénomène.

4.3.1. Rapport des stocks, rapport des flux et rapport des effectifs pertinents

Supposons qu'il existe deux populations distinctes $\mathcal{P}_1$ et $\mathcal{P}_2$ de type benchmark. La population $\mathcal{P}_1$ sera celle des reventes rapides et $\mathcal{P}_2$ celle des reventes lentes. Le paramètre λ de la fonction de survie étant une mesure directe de la vitesse de revente on aura donc $\lambda_1 > \lambda_2$. Les stocks seront notés S_1 et S_2 , les flux K_1 et K_2, sans préciser la date puisque dans ce modèle ces grandeurs ne dépendent pas du temps. Le rapport des flux est lié au rapport des stocks par la relation suivante :

$$S_2 / S_1 = (K_2 / (1 - e^{-\lambda_2})) / (K_1 / (1 - e^{-\lambda_1})) = (K_2 / K_1) * \left((1 - e^{-\lambda_1}) /(1 - e^{-\lambda_2})\right)$$

et en notant $V = (1 - e^{-\lambda_1}) /(1 - e^{-\lambda_2})$, il vient : $$S_2 / S_1 = V * (K_2 / K_1) \quad (43)$$

Avec l'approximation usuelle[42] on a $V \approx \lambda_1 / \lambda_2$. Le rapport des stocks et le rapport des flux sont donc liés par l'intermédiaire du rapport des vitesses.

[42] $e^x \approx 1 + x$ pour $x \approx 0$

Pour illustrer le propos, prenons par exemple T = 100, S_1 = 10 000, S_2 = 100 000, $\lambda_1 = 0{,}5$ et $\lambda_2 = 0{,}1$ (on a alors $V \approx 4{,}13$). L'espérance de détention[43] pour un bien de $\mathcal{P}_1$ est de 2 ans, de 10 ans pour un bien de type $\mathcal{P}_2$. Les flux[44] valent $K_1 \approx 3935$, $K_2 \approx 9516$. A chaque date, on observe ainsi 2,4 fois plus de biens $\mathcal{P}_2$ que de biens $\mathcal{P}_1$ échangés sur le marché ($K_2/K_1 = 2{,}4$), alors qu'ils sont en fait 10 fois plus nombreux dans le parc immobilier ($S_2/S_1 = 10$).

Le nombre de couples pertinents, pour un intervalle unitaire [t,t+1], pour chacune des populations (i = 1,2) est :

$$n_i^t = (K_i / (1 - \alpha_i)) * d_i (T- t)\, d_i(t + 1) = S_i\, d_i (T- t)\, d_i(t + 1)$$

d'où : $$n_2^t / n_1^t = (S_2 / S_1) * [\ (d_2(T- t)\, d_2(t + 1)) / (d_1(T- t)\, d_1(t + 1))] \qquad (44)$$

avec $d_i(k) = 1 - \alpha_i^k$ donnant le pourcentage de reventes réalisées après k unités de temps pour les biens de type $\mathcal{P}_i$.

Notons par $f(t) = f_2(t) / f_1(t) = (d_2(T- t)\, d_2(t + 1)) / (d_1(T- t)\, d_1(t + 1))$ le facteur qui modifie le rapport des stocks S_2 / S_1, dans le calcul de n_2^t / n_1^t. Ce coefficient mesure la divergence entre les parcs immobiliers réels et les observations recueillies dans l'échantillon. L'annexe 22 étudie les variations de f sur [0,T – 1]. Posons m = (T+1)/2 et a = (T – 1)/2 = m – 1, on a alors :

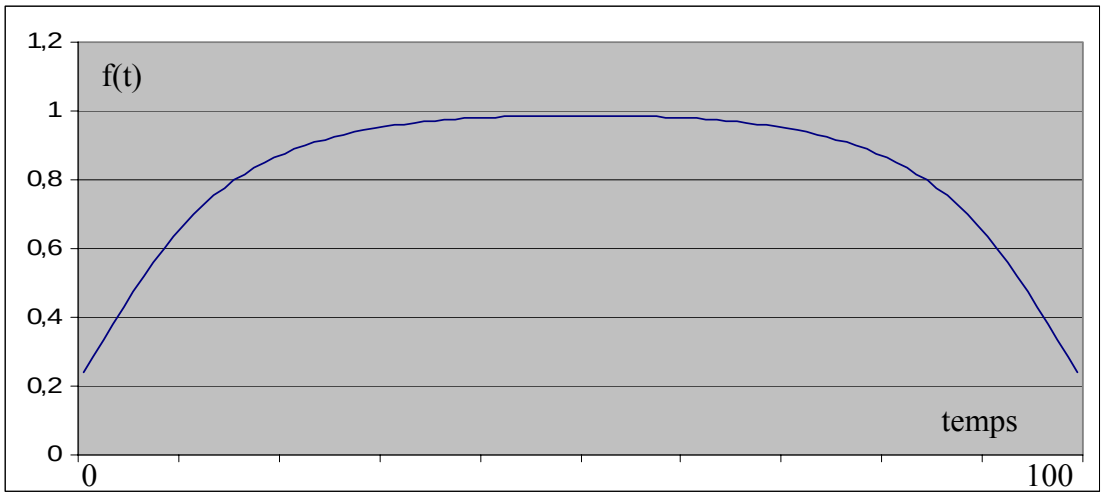
- Pour $\lambda_1 > \lambda_2$, la fonction f est croissante sur [0,a], atteint son maximum en a puis décroît.
- f est symétrique par rapport à l'axe vertical d'équation x = a
- f est majorée par 1 et son maximum vaut $M(\lambda_1, \lambda_2, T) = [d_2(m)/d_1(m)]^2$

[43] $X \sim \mathcal{E}(\lambda) \Rightarrow E(X) = 1 / \lambda$

[44] On rappelle que $S = K / (1 - \alpha)$ où $\alpha = e^{-\lambda}$

En reprenant les valeurs de l'exemple numérique ci-dessus, la figure 7 donne la représentation graphique de f.

Figure 7 : Facteur de modification du rapport des stocks dans le calcul de n_2^t / n_1^t
T = 100, $\lambda_1 = 0{,}5$ et $\lambda_2 = 0{,}1$



4.3.2. Commentaires

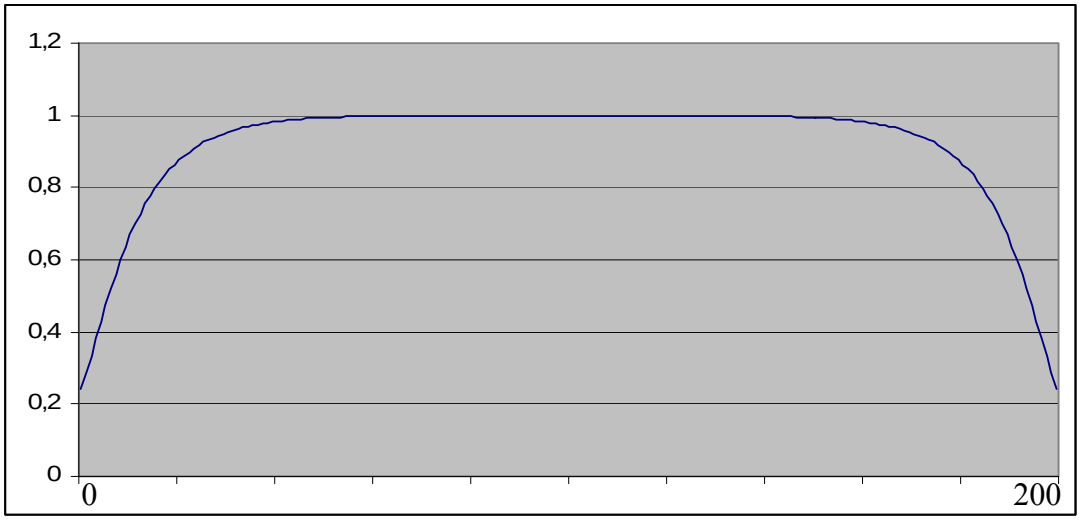
Quand f(t) est voisin de 1, les effectifs pertinents pour [t,t+1], de type 1 et de type 2, sont représentatifs des populations du stock dans le sens où leur proportion relative est voisine de celle constatée pour les stocks. Par contre, une valeur de f(t) sensiblement inférieur à 1 indiquera que la population de type $\mathcal{P}_2$ est sous-représentée ou, de manière équivalente, que $\mathcal{P}_1$ est surreprésentée.

Dans la figure 7, on constate que le problème du surnombre des « flips » se manifeste aux bords de l'intervalle [0,T]. Le facteur f(t) est supérieur à 0,95 pour les valeurs de t entre 30 et 70. Cela signifie, en d'autres termes, qu'un problème de représentativité significatif[45] se pose pour l'intervalle [0,30] et symétriquement pour [70,100] ; ces deux parties constituant 60% de l'intervalle [0,100]. Si l'on garde les

[45] $f(t) < 0{,}95$

mêmes paramètres, mais que l'on étend la fenêtre d'observation à l'intervalle [0,200] (figure 8), la situation s'améliore. On a ainsi f(t) > 0,95 pour tous les t de l'intervalle [30,170] et le problème de surreprésentation ne se pose plus alors que pour 30% de la fenêtre d'observation.

Figure 8 : Facteur de modification du rapport des stocks dans le calcul de n_2^t / n_1^t
T = 200, $\lambda_1 = 0,5$ et $\lambda_2 = 0,1$



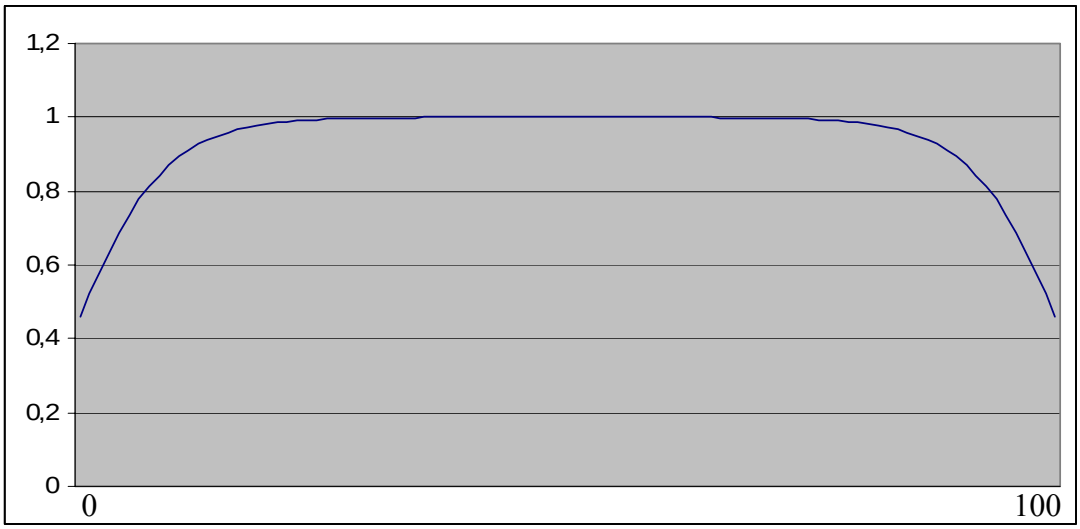
Ce problème n'est pas d'ordre économique, il est fondamentalement lié à la technique d'observation. Le voir comme un problème émanant de l'objet étudié reviendrait alors à considérer que les 2 stocks S_1 et S_2 ne sont pas dans de bonnes proportions. Or ce rapport est une donnée du marché immobilier, c'est donc à la technique indicielle de s'y adapter. La surreprésentation est en fait une conséquence mécanique de la méthode d'observation qui consiste à faire une césure à 0 et une autre à T. Pour prendre une métaphore on peut affirmer, dans l'absolu, que la surface de l'eau est parfaitement horizontale. Si cette eau se trouve dans un verre sa surface sera bien sûr pour l'essentiel toujours plane, mais près des parois celle-ci aura légèrement tendance à s'incurver vers le haut en raison de l'existence des bords. Dans l'exemple des deux populations il en est de même. Près des extrémités de [0,T],

$\mathcal{P}_1$ a tendance à être à un niveau plus élevé que dans l'absolu, c'est-à-dire sans bords temporels.

D'une manière générale, ces effets de bords font partie d'un concept plus global mit en exergue par Heisenberg (1932) : le principe d'incertitude. Observer n'est pas neutre et cet acte peut parfois modifier le phénomène sous le microscope.

Comme il a déjà été mentionné la longueur de la fenêtre d'observation est un facteur déterminant de la surreprésentation des « flips ». L'autre paramètre à prendre en compte est le rapport des vitesses. Si on reprend les paramètres de la figure 7, avec cette fois une valeur de 0,2 pour λ_2, le problème persiste mais il est atténué (cf. figure 9). Seulement 28% de l'intervalle [0,100] est associé à un coefficient f(t) inférieur à 0,95.

Figure 9 : Facteur de modification du rapport des stocks dans le calcul de n_2^t / n_1^t
T = 100, $\lambda_1 = 0{,}5$ et $\lambda_2 = 0{,}2$



Il semble qu'il soit assez difficile de se débarrasser complètement de l'effet de bord. Si les paramètres sont T = 100, $\lambda_1 = 0.5$ et $\lambda_2 = 0.4$, c'est-à-dire si les deux populations ont des comportements très voisins, il restera encore 8% de l'intervalle [0,100] où f(t) sera inférieur à 0.95.

En résumé, le problème de la surreprésentation d'une population dans l'échantillon se manifeste aux bords de l'intervalle d'étude. Il est d'autant plus marqué que la fenêtre d'observation est restreinte et que les vitesses de revente sont différentes. Le paragraphe suivant va essayer de quantifier plus précisément l'impact d'un surnombre sur les valeurs indicielles.

4.4. Population dominante, population dominée

Que se passe-t-il concrètement lorsque l'on estime l'indice sur un échantillon constitué de deux populations distinctes ? Quels sont les résultats en termes de tendance et de volatilité? Retrouve-t-on les caractéristiques des deux sous-échantillons, ou l'un des deux domine-t-il ? Ce paragraphe présente quelques éléments de réponse.

4.4.1. Principe de simulation et indicateurs agrégés

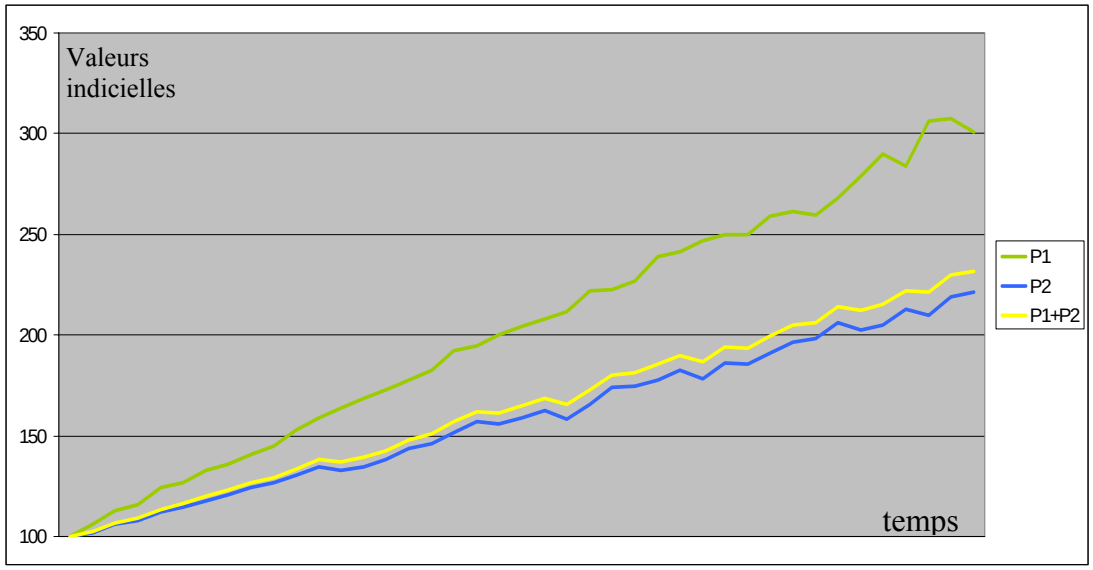
En utilisant le simulateur de données on génère deux populations $\mathcal{P}_1$ et $\mathcal{P}_2$ pour lesquelles $\sigma = 0{,}05$ et $\Theta = 10$. La courbe des "vrais prix" pour $\mathcal{P}_1$ est une fonction affine croissante, dont la variation sur chaque intervalle élémentaire vaut +5. Elle croit régulièrement de 100 (t = 0) à 300 (t = 40). Pour $\mathcal{P}_2$ la hausse des prix est plus lente : +3 à chaque intervalle. On évolue ainsi de 100 (t = 0) à 220 (t = 40).
Pour la simulation des reventes à l'intérieur de ces deux populations nous adoptons une modélisation de quasi-benchmark exponentiel. On fixe par conséquent les niveaux des stocks et des vitesses de revente, (S_1, λ_1) et (S_2, λ_2), et donc aussi indirectement les niveaux des flux constants échangés sur le marché à chaque date

(K_1, K_2). Les reventes sont calculées en appliquant aux effectifs encore en vie, à chaque date et pour chaque population, le taux de survie suivant[46]:

$$\text{Taux} = 1 - (\lambda_i + \varepsilon_i) \qquad i = 1,2 \text{ et } \varepsilon_i \text{ suit une loi uniforme sur } [-0{,}005 ; 0{,}005]$$

On calcule dans un premier temps un indice pour chacune des populations ($ind_{\mathcal{P}1}(t)$ et $ind_{\mathcal{P}2}(t)$ pour t = 0,…,T) puis en regroupant les deux échantillons on estime un indice global ($ind_{\mathcal{P}1+\mathcal{P}2}(t)$ pour t = 0,…,T). Ce dernier sera bien sûr situé "entre" les deux premiers indices. La figure 10 illustre graphiquement le type de résultat que l'on peut obtenir.

Figure 10 : Exemple de calcul d'indices pour $\mathcal{P}_1$, $\mathcal{P}_2$ et $\mathcal{P}_1 + \mathcal{P}_2$



Pour analyser les résultats, on définit pour chaque simulation deux types d'indicateurs. Le premier consiste à calculer pour toutes les dates le pourcentage de déviation[47] de la valeur indicielle $ind_{\mathcal{P}_i}(t)$ par rapport à $ind_{\mathcal{P}1+\mathcal{P}2}(t)$ et à faire la

[46] Le ε_i intervient ici pour perturber la structure trop régulière du benchmark exponentiel.
[47] $100x(ind_{\mathcal{P}_i}(t)/ind_{\mathcal{P}1+\mathcal{P}2}(t) - 1)$

moyenne de ces écarts sur [0,40]. On notera ECART($\mathcal{P}i$ / $\mathcal{P}1 + \mathcal{P}2$) le résultat[48]. Cette grandeur servira à étudier la tendance qui domine dans l'indice. Une valeur faible de ECART($\mathcal{P}1$ / $\mathcal{P}1 + \mathcal{P}2$) indiquera par exemple que la courbe de $\mathcal{P}1$ est proche de celle de $\mathcal{P}1 + \mathcal{P}2$ ou, en d'autres termes, que la tendance de $\mathcal{P}1$ dirige celle de $\mathcal{P}1 + \mathcal{P}2$. Afin d'avoir des éléments de comparaison, si les indices $\mathcal{P}1$ et $\mathcal{P}2$ reproduisent parfaitement leur courbe des vrais prix, et que la courbe de $\mathcal{P}1 + \mathcal{P}2$ est équidistante de $\mathcal{P}1$ et de $\mathcal{P}2$, on obtient les valeurs suivantes :

ECART($\mathcal{P}1$ / $\mathcal{P}1 + \mathcal{P}2$) = 10,26 ECART($\mathcal{P}2$ / $\mathcal{P}1 + \mathcal{P}2$) = - 10,26

Le second type d'indicateur, destiné à capturer la volatilité dominante dans l'indice agrégé, consistera simplement à calculer les coefficients de corrélation entre les taux de rentabilité monopériodiques pour $\mathcal{P}1$, ou $\mathcal{P}2$, et ceux de l'indice $\mathcal{P}1 + \mathcal{P}2$; on les notera CORR($\mathcal{P}i$ / $\mathcal{P}1+\mathcal{P}2$). Plus l'indicateur sera proche de 1, plus la volatilité de la population $\mathcal{P}i$ associée dirigera celle de l'indice global.

4.4.2. Les premiers résultats

Les deux types d'indicateurs définis précédemment dépendent des échantillons générés, ils présentent donc une certaine volatilité. Afin d'obtenir des valeurs raisonnablement fiables les résultats présentés ci-dessous, tableau 4 et tableau 5, correspondent aux valeurs moyennes obtenues à partir de 20 simulations[49]. Fixons les caractéristiques de la première population à $S_1 = 100000$ et $\lambda_1 = 0,05$ (on rappelle pour mémoire que l'espérance d'une loi exponentielle de paramètre λ est $1 / \lambda$, ce choix signifie donc que les biens de la population $\mathcal{P}1$ sont conservés en moyenne pendant 20 ans) et étudions les valeurs des différents indicateurs, en fonction de S_2 et λ_2.

[48] On mesure en fait la distance entre les deux courbes
[49] 20 itérations suffisent pour obtenir un bon aperçu

Tableau 4 : ECART($\mathcal{P}1 / \mathcal{P}1 + \mathcal{P}2$) et ECART($\mathcal{P}2 / \mathcal{P}1 + \mathcal{P}2$) en fonction de S_2 et λ_2
(la première valeur donne ECART($\mathcal{P}1 / \mathcal{P}1+\mathcal{P}2$), la deuxième ECART($\mathcal{P}2 / \mathcal{P}1+\mathcal{P}2$))

S_2 / λ_2	**20 000**	**50 000**	**80 000**	**100 000 (= S_1)**	**120 000**	**150 000**	**200 000**
0,01	0 - 18	1 - 18	1 - 17	2 - 17	2 - 17	2 - 17	3 - 16
0,02	1 - 17	3 - 16	4 - 16	5 - 15	5 - 15	6 - 14	8 - 13
0,04	3 - 16	6 - 14	8 - 12	9 - 11	10 - 10	12 - 9	13 - 8
0,05 (= λ_1)	3 - 15	7 - 13	10 - 11	10 - 10	12 - 9	13 - 8	15 - 7
0,1	6 - 14	11 - 10	13 - 8	14 - 7	15 - 6	16 - 5	17 - 4
0,2	7 - 13	13 - 9	15 - 7	15 - 6	17 - 6	16 - 4	17 - 4
0,3	7 - 13	12 - 9	14 - 7	16 - 7	17 - 6	16 - 5	19 - 4

Tableau 5 : CORR ($\mathcal{P}1 / \mathcal{P}1 + \mathcal{P}2$) et CORR ($\mathcal{P}2 / \mathcal{P}1 + \mathcal{P}2$) en fonction de S_2 et λ_2
(la première valeur donne CORR ($\mathcal{P}1 / \mathcal{P}1+\mathcal{P}2$), la deuxième CORR($\mathcal{P}2 / \mathcal{P}1+\mathcal{P}2$))

S_2 / λ_2	**20 000**	**50 000**	**80 000**	**100 000 (= S_1)**	**120 000**	**150 000**	**200 000**
0,01	1 0.07	1 0.21	1 0.15	1 0.07	1 0.16	1 0.22	0.99 0.15
0,02	1 0.10	1 0.18	0.99 0.23	0.99 0.27	0.98 0.29	0.97 0.33	0.94 0.38
0,04	0.99 0.21	0.95 0.46	0.91 0.51	0.86 0.58	0.80 0.66	0.79 0.72	0.68 0.79
0,05 (= λ_1)	0.98 0.27	0.91 0.48	0.81 0.67	0.75 0.70	0.69 0.76	0.63 0.83	0.52 0.87
0,1	0.85 0.60	0.61 0.86	0.48 0.94	0.4 0.95	0.32 0.97	0.27 0.98	0.23 0.99
0,2	0.51 0.90	0.27 0.97	0.21 0.99	0.15 0.99	0.13 1	0.14 1	0.07 1
0,3	0.35 0.97	0.19 0.99	0.14 1	0.16 1	0,02 1	0.14 1	0.10 1

Lorsque les deux populations ont des caractéristiques identiques ($S_1 = S_2 = 100000$ et $\lambda_1 = \lambda_2 = 0{,}05$), la tendance de l'indice agrégé correspond à la « moyenne » des deux sous-tendances et la volatilité globale reflète en d'égales proportions les deux sous-volatilités (CORR ($\mathcal{P}1 / \mathcal{P}1 + \mathcal{P}2$) = 0.75 , CORR($\mathcal{P}2 / \mathcal{P}1 + \mathcal{P}2$) = 0.70).

Le tableau 4 des indicateurs de tendance montre que ECART($\mathcal{P}1 / \mathcal{P}1 + \mathcal{P}2$) augmente avec S_2 et λ_2. Dans le même temps, ECART($\mathcal{P}2 / \mathcal{P}1 + \mathcal{P}2$) décroît en valeurs absolues. L'indicateur de volatilité CORR ($\mathcal{P}1 / \mathcal{P}1 + \mathcal{P}2$) évolue à l'opposé de S_2 et λ_2, tandis que CORR($\mathcal{P}2 / \mathcal{P}1 + \mathcal{P}2$) varie dans le même sens (tableau 5). Ces résultats sont en accord avec l'intuition que l'on peut en avoir si l'on remarque que la part de la population $\mathcal{P}2$ dans l'échantillon s'accroît avec le stock S_2, et le taux de rotation des biens, capturé par λ_2. Il est donc normal de constater que dans cette situation les caractéristiques de $\mathcal{P}2$ ressortent plus que celles de $\mathcal{P}1$.

L'impact sur la tendance commune est réel, mais l'élément le plus notable dans cette simulation concerne la volatilité. Les coefficients de corrélation prennent en effet rapidement des valeurs proches de 1. Il semble donc que la volatilité capturée par un indice global soit représentative d'une seule des deux populations quand l'échantillon est déséquilibré. On parlera dans ce cas de population dominante et de population dominée. En termes d'investissement ce résultat n'est pas sans conséquences. La perception du risque moyen de l'immobilier pouvant être altérée si le comportement de l'indice est dirigé par un seul type de bien, baser une allocation d'actifs sur un tel indicateur c'est alors courir le risque de la sous-optimalité.

A titre d'exemple deux cases ont été grisées dans les tableaux 4 et 5. Elles mettent en évidence ce qui se produit lorsque l'on passe d'un stock homogène de 200000 biens avec $\lambda_1 = \lambda_2 = 0{,}05$ (détention moyenne 20 ans) à un échantillon où 100000 de ces biens deviennent des "flips" : $\lambda_2 = 0{,}2$ (détention moyenne 5 ans). Le décalage de l'indice agrégé vers l'indice $\mathcal{P}2$ est sensible mais il reste, en termes de

tendance, raisonnable ; la population $\mathcal{P}2$ ne domine pas radicalement $\mathcal{P}1$ sur ce point. Par contre en termes de volatilité, on passe de 0,80 à 0,19 pour $\mathcal{P}1$ et de 0,74 à 0,99 pour $\mathcal{P}2$. Le risque retranscrit par l'indice global est donc presque uniquement celui des "starter homes"[50], $\mathcal{P}2$ domine $\mathcal{P}1$ dans l'expression du risque.

4.4.3. Désagrégation des indicateurs de volatilité

Les figures 7, 8 et 9 indiquent que le rapport des effectifs pertinents n_2^t / n_1^t s'écarte du rapport des stocks principalement aux bords de l'intervalle (lorsque les vitesses sont différentes). Les phénomènes de domination ne devraient donc pas être uniformes sur toute la période d'estimation. Pour tester cette affirmation nous désagrègons les indicateurs CORR($\mathcal{P}i$ / $\mathcal{P}1+\mathcal{P}2$) en trois sous-indicateurs[51] : $CORR_d$($\mathcal{P}i$ / $\mathcal{P}1+\mathcal{P}2$), $CORR_m$($\mathcal{P}i$ / $\mathcal{P}1+\mathcal{P}2$), $CORR_f$($\mathcal{P}i$ / $\mathcal{P}1+\mathcal{P}2$). Le premier mesure le coefficient de corrélation des taux de rentabilité entre $\mathcal{P}i$ et $\mathcal{P}1 + \mathcal{P}2$ sur l'intervalle [0,5], le deuxième sur [18,23] et le troisième sur [36, 40]. Ces mesures étant plus volatiles que CORR($\mathcal{P}i$ / $\mathcal{P}1+\mathcal{P}2$), on réalise pour chaque cas 500 simulations. Les résultats sont présentés dans le tableau 6 (le nombre de valeurs testées pour S_2 et λ_2 est plus restreint que précédemment pour des raisons de temps de calcul)

Pour interpréter les résultats on rappelle que le lien entre le rapport des stocks et le rapport des effectifs pertinents s'écrit : $n_2^t / n_1^t = (S_2 / S_1) * f(t)$. Pour $\lambda_1 > \lambda_2$ la fonction f est croissante sur [0,a], atteint son maximum en a puis décroît. Elle est symétrique par rapport à l'axe vertical d'équation x = a et majorée par 1. Pour $\lambda_1 < \lambda_2$, la situation est inversée, la courbe de f n'est plus en forme de cloche mais en forme de creux, cf figure 11.

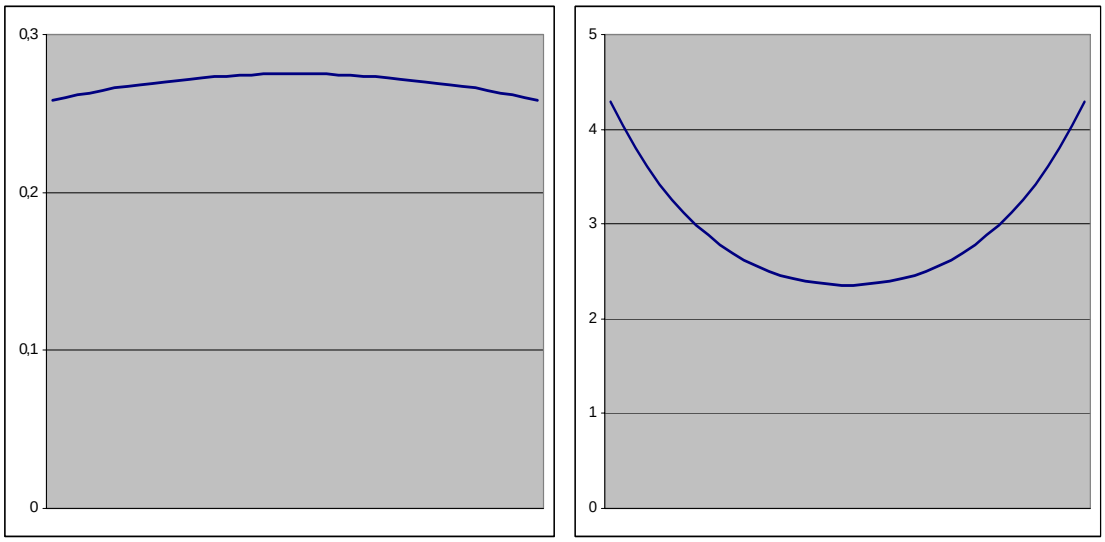
[50] Flips ou logements de primo-accédant : studio-T1-T2
[51] "d" pour début, "m" pour milieu et "f" pour fin

Tableau 6 : Indicateurs CORR(*Pi* / *P1*+*P2*) désagrégés
(sous chaque valeur agrégée on trouve respectivement $CORR_d$, $CORR_m$, $CORR_f$; la première ligne concerne *P1*, la deuxième *P2*)

λ_2 \ S_2	**50 000**	**100 000 (= S_1)**	**150 000**
0,02	1 (0.99 0.99 0.99) 0.18 (0.09 0.11 0.05)	0.99 (0.98 0.98 0.98) 0.27 (0.17 0.16 0.11)	0.97 (0.95 0.95 0.95) 0.33 (0.22 0.25 0.20)
0,05 (= λ_1)	0.91 (0.87 0.86 0.87) 0.48 (0.39 0.40 0.36)	0.75 (0.64 0.63 0.66) 0.70 (0.63 0.66 0.65)	0.63 (0.49 0.49 0.53) 0.83 (0.81 0.80 0.78)
0,2	0.27 (0.18 0.18 0.16) 0.97 (0.98 0.97 0.98)	0.15 (0.05 0.08 0.08) 0.99 (0.99 0.99 0.99)	0.14 (0.07 0.02 0.08) 1 (1.00 1.00 1.00)

Figure 11 : Représentation graphique de la fonction f sur [0,40]

$\lambda_1 = 0{,}05$ et $\lambda_2 = 0{,}02$ $\lambda_1 = 0{,}05$ et $\lambda_2 = 0{,}2$



Dans la cellule gris clair et la cellule gris foncé du tableau 6, les deux facteurs λ et S agissent dans le même sens[52]. En haut à gauche, la population $\mathcal{P}2$ est sous-pondérée, tandis qu'à l'opposée du tableau, la population $\mathcal{P}1$ est sous-représentée (en bas à droite). Dans le premier cas les effectifs de type $\mathcal{P}2$ sont moins nombreux, en termes relatifs, aux bords de l'intervalle comme on l'a vu dans le paragraphe 4.3. En conséquence, les indicateurs désagrégés de volatilité exhibent logiquement une courbe en bosse[53] traduisant une meilleure capture de la volatilité de type $\mathcal{P}2$ au centre de l'intervalle. Dans le deuxième cas la situation s'inverse, la population de type $\mathcal{P}1$ étant relativement à $\mathcal{P}2$ mieux représentée aux bords, on obtient des indicateurs en creux[54] (cf deuxième graphique de la figure 11). Ici l'indice global capture mieux le risque de $\mathcal{P}1$ au début et à la fin de l'intervalle qu'au centre. Pour les autres cas de figures, les résultats paraissent plus difficiles à interpréter. Pour certains d'entre eux une asymétrie des volatilités semble être à l'œuvre[55], ces phénomènes pourraient sans doute faire l'objet d'approfondissements.

5. Conclusion

Le formalisme développé dans le chapitre 1 confirme son efficacité pour appréhender les problèmes et les caractéristiques complexes du RSI. En s'appuyant sur la décomposition de l'indice en ses constituants élémentaires et en précisant quelques hypothèses économiques de la modélisation, nous avons ainsi pu établir des formules simples qui rendent compte de la volatilité de l'indice ($\mathcal{V}(R) = \sigma_G^2 \hat{I}^{-1}$) et de sa réversibilité ($\hat{I}(T_2)R(T_2) = \hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)$). La quantification de cette dernière est un enjeu d'importance pour l'introduction de produits dérivés sur

[52] Pour λ_1 fixé, la surreprésentation de $\mathcal{P}2$ augmente avec S_2 et λ_2
[53] Pour $\mathcal{P}2$: (0.09 0.11 0.05)
[54] Pour $\mathcal{P}1$: (0.07 0.02 0.08)
[55] Pour S_2 = 100 000 : quand λ_2 = 0,02 on a pour $\mathcal{P}2$ (0.17 0.16 0.11) et quand λ_2 = 0,2 on a pour $\mathcal{P}1$ (0.05 0.08 0.08)

indices immobiliers, car il semble qu'il soit nécessaire d'avoir une certaine connaissance de l'ampleur potentielle de ce phénomène pour pouvoir travailler correctement avec les dérivés indiciels. Enfin, en utilisant la modélisation du benchmark exponentiel, il a été possible d'analyser théoriquement et empiriquement le problème des deux populations. Les concepts d'effets de bords, de population dominante et de population dominée, se sont révélés centraux. Les problématiques non triviales présentées et analysées dans cette section ont en général trouvé leur aboutissement. Le prochain et dernier chapitre de cette thèse proposera d'autres thèmes de réflexion, dans une optique plus exploratoire, et il tentera de poser des jalons qui pourront servir au développement de futures recherches.

Chapitre 4

Vers une théorie des indices informationnels et autres prolongements

1. Introduction

En s'appuyant sur les résultats et les problématiques des sections précédentes, ce dernier chapitre explore des pistes pour de futures recherches. L'étude formelle de ces nouveaux thèmes sera initiée d'une manière approfondie et rigoureuse, mais elle restera toutefois un peu moins aboutie et un peu moins exhaustive que dans les chapitres précédents.

L'essentiel du travail de cette thèse a consisté à mettre au jour la structure informationnelle implicite aux indices de ventes répétées. Une fois que la définition de l'information est connue, l'estimation de l'indice se fait sans difficultés, en suivant l'algorithme de calcul présenté dans la figure 12 du chapitre 1. Le point névralgique du modèle concerne donc la définition de l'information. Dans les modèles et dans l'approche traditionnels, cette question n'est pas posée directement, on ne peut donc pas y répondre en pleine conscience. Or, comme calculer un indice dans un marché hétérogène revient en fait à répondre incontournablement à cette question, on est amené à quantifier cette grandeur sans vraiment s'en rendre compte. Des hypothèses informationnelles sont alors faites implicitement ce qui, dans une optique scientifique, est toujours délicat. Mais il y a plus grave encore. Le préjudice majeur de cette approche réside surtout dans la perte de flexibilité que l'on s'impose. On se prive en effet de la possibilité de définir l'information en fonction des problématiques économiques qui motivent le calcul de l'indice, et donc de la possibilité de rendre plus « parlant » l'indice. Il va bien sûr de soi que la définition de cette grandeur floue que l'on appelle l'information ne se fera pas sans difficultés, mais comme ce passage est incontournable, autant l'aborder en pleine connaissance de cause plutôt qu'avec un bandeau sur les yeux. Le premier paragraphe de ce chapitre consistera donc à explorer la façon dont on peut définir un indice à partir d'une quantification explicite, et la plus générale possible, de l'information.

Le deuxième paragraphe présentera des pistes d'approfondissements, plus mathématiques, où l'on étudiera la matrice $\hat{I}$ à l'aide de concepts géométriques. Comme mentionné ci-dessus la notion d'information est au cœur du concept

d'indice, l'étude mathématique de sa structure globale, que condense la matrice Î, est donc un objectif légitime. Il n'y aura pas d'applications financières directes à ces réflexions, mais l'analyse théorique des matrices informationnelles pourrait un jour devenir utile. Divers concepts seront introduits et discutés : contribution informationnelle moyenne, mesure de diffusion et de spécialisation de l'information, volumes informationnels …

Le dernier paragraphe sera d'une autre nature. Il devra être mis en perspective avec l'actualité économique du moment dans le domaine de la finance immobilière, à savoir la création de produits dérivés sur indices immobiliers. Comme leur nom l'indique ces actifs dérivent des indices, ils constituent donc un prolongement concret de ce travail sur le RSI. Leur évaluation ne pourra pas être raisonnablement réalisée sans avoir une très bonne connaissance de la dynamique et de la mécanique des actifs primitifs. On étudiera dans ce paragraphe, plus particulièrement, l'applicabilité de la théorie de l'arbitrage aux sous-jacents immobiliers et on tentera de poser quelques jalons pour l'évaluation des actifs contingents. Des difficultés non négligeables émergeront à cette occasion, la prudence et la rigueur devront donc être de mise pour l'étude de ce genre de problème.

2. Vers une théorie des indices informationnels

On tentera dans ce paragraphe de généraliser la notion d'indice. La problématique sera d'abord discutée et quelques réflexions préalables sur le procédé à employer seront développées. On procédera ensuite à l'énonciation des définitions qui seront illustrées par deux exemples numériques. Enfin, la dernière section traitera de la question la plus importante, à savoir la quantification de l'information ; les différentes options utilisées implicitement dans la littérature seront exposées.

2.1. La problématique

Lors de la reformulation de l'indice de ventes répétées, la notion de quantité d'information est apparue comme étant le concept central. Pour une vente répétée (t_i,t_j) elle se mesurait par une constante dans le modèle BMN, par $(\Theta + (j - i))^{-1}$ dans le modèle CS et par $(j - i)^{-1}$ dans le paragraphe 3 du chapitre 1. Ces définitions de la mesure d'information découlent directement de la spécification de la régression, et plus particulièrement des hypothèses faites sur le terme d'erreur ε : bruit blanc pour BMN, bruit blanc et marche aléatoire pour CS, marche aléatoire uniquement pour le dernier cas. Mais indépendamment de ces différences, les relations fondamentales entre les différentes briques élémentaires constituant le RSI sont maintenues entre ces trois modèles, comme on a pu le constater à plusieurs reprises. La relation $\hat{I}R = \eta P$ est par exemple toujours vraie. Une question naturelle se pose alors :

> *« Au lieu de définir indirectement la mesure d'information par les hypothèses de la régression, serait-il possible de construire l'indice en prenant explicitement les poids informationnels des ventes répétées comme des données initiales ? »*

Une telle approche autoriserait une plus grande flexibilité dans la définition de l'information, car celle-ci ne serait plus nécessairement constante ou de la forme $(j - i)^{-1}$ ou $(\Theta + (j - i))^{-1}$. On pourrait alors décider en pleine conscience, et selon les contextes et les problématiques des études économiques envisagées, de la représentativité d'un bien par rapport à l'échantillon global ou, en d'autres termes, de sa contribution à l'indice simplement en sous-pondérant ou en sur-pondérant sa quantité d'information. Si l'on prend l'exemple du marché immobilier de New York et un échantillon couvrant la période 1990-2002, il semble raisonnable de sur-pondérer les transactions réalisées après le 11 septembre 2001 car leur prix intègrent l'événement. Les ventes antérieures à cette date sont beaucoup moins pertinentes

pour juger de l'état du marché new-yorkais à la fin de l'année 2001 ; même si celles-ci ont été réalisées le 10 septembre.

La suite de ce paragraphe ne cherchera pas à déterminer une méthode pour décider correctement du niveau d'information contenue dans un bien particulier ; on supposera au contraire que ces nombres sont des données. Par contre, on s'attachera à trouver de nouvelles justifications théoriques pour le calcul de l'indice, car la relation-définition $\hat{I}\ R = \eta\ P$ ne pourra plus être considérée comme une conséquence du problème de minimisation. Quels seront alors ses fondements et son interprétation?

2.2. Comment généraliser ?

On reprendra ici les notations introduites dans les chapitres précédents sans les redéfinir explicitement, sauf si la situation l'impose. La réflexion développée dans ce paragraphe est exploratoire et les variables introduites sont provisoires. Il ne s'agit que d'une étape intermédiaire du processus d'abstraction, la version finale sera présentée dans le paragraphe suivant.

Si pour la $k^{ème}$ vente répétée dont l'achat a été réalisé à $t_i(k)$ au prix $p_{k,i}$ et la revente à $t_j(k)$ au prix $p_{k,j}$, on note par inf(k) la quantité d'information correspondante, pour la classe[1] de transaction (i,j) la quantité d'information totale se calculerait naturellement par $L_{i,j} = \sum_{k'}\text{inf}(k')$ (l'indice k' ne parcourant que les éléments de cette classe). inf(k) serait compris entre 0 (vente non informative) et $+\infty$ (vente parfaitement informative).

[1] Classe de transaction (i,j) : ensemble des ventes répétées pour lesquelles l'achat est à t_i et la revente à t_j.

A partir de la distribution des $\{L_{i,j}\}$, on obtiendrait par les sommations usuelles les quantités $I^{[t',t]}$, ainsi que la matrice $\hat{I}$. Pour définir la matrice η nous pourrions utiliser la troisième relation de la proposition 3 (chapitre 3, paragraphe 2.1). Celle-ci affirme que dans le cas traditionnel (CS ou BMN), la somme des éléments de la matrice $\hat{I}$ dans la ligne i est égale au $i^{ème}$ élément diagonal de la matrice η. Au lieu d'utiliser cette formule comme une propriété vérifiée par η on pourrait désormais lui donner le statut de définition.

La formule de ρ_t dans le chapitre 1 était : $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} G(j-i) r_{k'}^{(i,j)}$ (1)

Elle peut se réécrire : $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} (j - i) \left[(\Theta + (j - i))^{-1} r_{k'}^{(i,j)} \right]$

$n^t G(\zeta^t)$ étant le $t^{ème}$ élément diagonal de η et $(\Theta + (j-i))^{-1}$ la quantité d'information fournie par un bien dans le modèle classique, la généralisation naturelle consisterait à définir ρ_t par la formule :

$$\rho_t = (\eta(t,t))^{-1} \sum_{i \le t < j} \sum_{k'} \left[(j - i) \inf(k') \right] r_{k'}^{(i,j)} \qquad (2)$$

où : $$\eta(t,t) = \sum_{i \le t < j} \sum_{k'} \inf(k') (j - i) \qquad (2')$$

Puis, en appliquant la formule $R = \hat{I}^{-1} (\eta P)$, on pourrait enfin définir l'indice.

Il est toutefois possible et souhaitable d'aller un peu plus loin dans la reformulation des concepts pour deux raisons. En premier lieu choisir une telle définition pour ρ_t est un peu artificiel, il faut donc lui donner des soubassements et des justifications théoriques plus formels. D'autre part il serait assez intéressant de

construire un modèle dans lequel les quantités inf(k') et $r_{k'}^{(i,j)}$ seraient autorisées à varier au cours de la période de détention du bien.

2.3. Définition théorique d'un indice informationnel

Le concept pivot pour définir un indice informationnel sera la moyenne des taux moyens ρ_t. Nous le définirons tout d'abord, de la façon la plus générale possible, dans le premier paragraphe pour un échantillon réel (on parlera de taux moyen réel). Puis, en substituant aux rendements réalisés par les biens de l'échantillon, une série de rendements universels on établira la formule définissant le taux moyen indiciel. Ces deux concepts permettront alors de donner une définition rigoureuse au concept d'indice informationnel.

2.3.1. Réécriture de ρ_t

Dans l'écriture précédente de ρ_t, pour la k'^{ème} vente répétée de la classe (i,j) on considère implicitement que le taux moyen $r_{k'}^{(i,j)} = \ln(p_{k,j} / p_{k,i}) / (j - i)$ s'applique sur chacun des sous-intervalles unitaires de la période $[t_i, t_j]$. Si on ne connaît que le prix d'achat et le prix de revente du bien et que l'on ne dispose pas d'informations sur les fluctuations intermédiaires du prix, cette hypothèse est très probablement la meilleure à poser. Le terme $(j - i) r_{k'}^{(i,j)}$ qui apparaît dans la formule (2) peut, en effet, se réécrire comme $r_{k'}^{(i,j)}([i,i+1]) + r_{k'}^{(i,j)}([i+1,i+2]) + \ldots + r_{k'}^{(i,j)}([j - 1, j])$ où ces différents taux intermédiaires sont simplement égaux au taux unitaire constant $r_{k'}^{(i,j)}$ pour la k'^{ème} vente répétée de la classe (i,j). ρ_t s'écrit donc sous une forme développée :

$$\rho_t = (\eta(t,t))^{-1} \sum_{i \le t < j} \sum_{k'} \sum_{s=i,\ldots,j-1} \inf(k') \, r_{k'}^{(i,j)}([s,s+1]) \qquad (3)$$

La pondération informationnelle de chacun des intervalles [i,i+1], [i+1,i+2],…, [j - 1, j] est identique et égale à inf(k'), pour un bien k'. Dans ce modèle il n'y a en effet pas de raisons particulières pour valoriser différemment l'approximation $r_{k'}^{(i,j)} = \ln(p_{k,j}/p_{k,i})/(j-i)$ entre les différents intervalles unitaires de la période de détention.

Ces deux remarques suggèrent un cadre plus global pour les indices. Supposons que l'on dispose pour le bien k d'une information qui ne se limite pas à l'observation d'un prix initial et d'un prix final. Le bien peut par exemple avoir été vendu plus de deux fois ou des expertises de sa valeur ont pu être réalisées au cours de la période de détention. Si l'on autorise les quantités inf(k') et $r_{k'}^{(i,j)}$ à varier entre i et j, il serait alors possible d'inclure ces renseignements dans le calcul de l'indice. On pourrait même aller jusqu'à imaginer une situation où ces valeurs seraient différentes sur chacun des sous-intervalles élémentaires.

Cette approche présente plusieurs avantages. Les biens pour lesquels le nombre de ventes est supérieur strictement à 2 sont problématiques pour le RSI car on se retrouve alors dans une situation d'autocorrélation des résidus (cf. Shiller (1991)). Une écriture informationnelle directe de l'indice pourrait être un moyen de contourner ce problème. D'autre part, l'incorporation de données d'expertise dans le calcul du RSI rapprocherait cet indice des méthodes hybrides. L'estimation se ferait alors sur des prix réels et des prix estimés, augmentant ainsi l'informativité de l'indice. Une valeur d'expertise est bien sûr moins fiable qu'un prix de transaction, mais cette difficulté pourrait se gérer aisément avec l'approche informationnelle ; il suffirait pour cela de réduire correctement les pondérations des intervalles de temps concernés.

En toute généralité, on définira donc ρ_t par :

$$\rho_t = (\eta(t,t))^{-1} \sum_{i \le t < j} \sum_{k' \varepsilon\, C(i,j)} \sum_{s = i,\ldots,j-1} \mathrm{inf}_{k'}(s)\, r_{k'}(s) \qquad (4)$$

avec : $$\eta(t,t) = \sum_{i \le t < j} \sum_{k' \varepsilon\, C(i,j)} \sum_{s = i,\ldots,j-1} \mathrm{inf}_{k'}(s) \qquad (4')$$

Les quantités $inf_{k'}(s)$ et $r_{k'}(s)$ ne seront plus ici nécessairement égales sur [i,j]. Les notations ayant légèrement varié nous les reprécisons explicitement. La première sommation parcourt l'ensemble des classes[2] de ventes répétées $C(i,j)$ avec $i \leq t < j$. Dans une classe $C(i,j)$ l'indice k' parcourt les différents éléments (deuxième sommation). Enfin, pour un bien k dans $C(i,j)$, $r_{k'}(s)$ est la valeur fournie par les données pour le taux de croissance de ce bien particulier au cours de l'intervalle [s,s+1] et $inf_{k'}(s)$ est la pondération informationnelle associée.

2.3.2. Taux moyen réel et taux moyen indiciel

On appellera vecteur réel des taux moyens, le vecteur de dimension T obtenu en appliquant la formule 4 pour t = 0,…,T-1 ; il sera noté $P_{réel}$. Supposons maintenant que pour chaque intervalle de temps unitaire les taux $r_{k'}(s)$ ne dépendent pas des biens particuliers mais simplement de l'intervalle considéré, on note par $(r_0, r_1, \ldots, r_{T-1})$ cette série universelle. Tout en conservant les pondérations informationnelles spécifiques aux biens, voyons ce que deviennent les quantités ρ_t sous cette hypothèse. Le vecteur ainsi obtenu sera appelé, pour des raisons qui apparaîtront par la suite, le vecteur indiciel des taux moyens. On le notera P_{ind}.

Initialement nous avons : $\rho_t = (\eta(t,t))^{-1} \sum_{i \leq t < j} \sum_{k' \varepsilon\, C(i,j)} \sum_{s = i,\ldots,j-1} inf_{k'}(s)\, r_{k'}(s)$

En remplaçant $r_{k'}(s)$ par son équivalent universel r_s il vient :

$$\rho_t = (\eta(t,t))^{-1} \sum_{i \leq t < j} \sum_{k' \varepsilon\, C(i,j)} \sum_{s = i,\ldots,j-1} inf_{k'}(s)\, r_s$$

$$\rho_t = (\eta(t,t))^{-1} \sum_{i \leq t < j} \sum_{s = i,\ldots,j-1} \left[\sum_{k' \varepsilon\, C(i,j)} inf_{k'}(s) \right] r_s$$

[2] $C(i,j)$: achat à t_i, revente à t_j

Pour poursuivre le calcul il faut remarquer que la variabilité de l'information $inf_{k'}(s)$ entre i et j, pour les ventes répétées k' de *C(i,j)*, amène à définir une famille de $L_{i,j}(s)$ (pour s = i,…, j-1), au lieu d'utiliser une seule et même mesure $L_{i,j}$, comme dans le modèle classique. Précisément nous noterons :

$$L_{i,j}(s) = \sum_{k' \varepsilon\, C(i,j)} inf_{k'}(s) \qquad \text{pour } s = i, \dots, j-1 \qquad (5)$$

On a alors : $$\rho_t = (\eta(t,t))^{-1} \sum_{i \le t < j} \sum_{s = i,\dots,j-1} L_{i,j}(s)\, r_s \qquad (6)$$

Ce type de somme, $\sum_{i \le t < j} \sum_{s = i,\dots,j-1} L_{i,j}(s)\, r_s$, a déjà été rencontré et transformé dans les démonstrations précédentes pour des $L_{i,j}$ constants (cf. annexe 2 par exemple). Pour la rendre plus interprétable nous la réordonnons en fonction des taux r_s, en réutilisant la même technique de calcul.

Ainsi pour $t' \le t$, le nombre de $r_{t'}$ est :

$$(L_{0,t+1}(t') + \dots + L_{0,T}(t')) + (L_{1,t+1}(t') + \dots + L_{1,T}(t')) + \dots + (L_{t',t+1}(t') + \dots + L_{t',T}(t')) \qquad (7)$$

et pour $t' > t$:

$$(L_{0,T}(t') + \dots + L_{t,T}(t')) + (L_{0,T-1}(t') + \dots + L_{t,T-1}(t')) + \dots + (L_{0,t'+1}(t') + \dots + L_{t,t'+1}(t')) \qquad (7')$$

Dans le premier cas les indices des $L_{i,j}(t')$ indiquent que l'on considère toutes les ventes répétées dont la période de détention inclut [t',t+1], dans le second celles dont la période inclut [t,t'+1]. En d'autres termes et sans distinguer les deux cas, nous travaillons avec la population des ventes répétées pertinentes pour [t,t+1] et on ne conserve que celles dont la période de détention inclut [t',t'+1] pour déterminer le coefficient devant chaque $r_{t'}$.

Pour chaque type de $L_{i,j}$ apparaissant dans ces sommes, la valeur de la série des $L_{i,j}(s)$ retenue est simplement celle qui correspond au taux $r_{t'}$, c'est-à-dire à l'intervalle

[t',t'+1] : $L_{i,j}(t')$. Pour les ventes répétées sélectionnées, on additionne donc les quantités d'information qu'elles fournissent pour l'intervalle [t',t'+1] pour obtenir le nombre de $r_{t'}$ dans l'écriture de ρ_t.

D'une manière plus synthétique, le vecteur $P_{ind} = (\rho_0, \rho_1, \ldots, \rho_{T-1})'$ des taux moyens que l'on obtient lorsque l'on remplace les taux $\{r_{k'}(s)\}$ par le vecteur des taux universels $R = (r_0, r_1, \ldots, r_{T-1})'$ peut s'exprimer par une formule matricielle utilisant une matrice $\hat{I}$ et une matrice η. Les composants de la $i^{ème}$ ligne de $\hat{I}$ se calcule à partir de la population $\mathcal{P}op_i$ des ventes répétées pertinentes pour l'intervalle [i,i+1]. Le $j^{ème}$ élément de cette ligne donne la quantité d'information fournie par ces données pour l'intervalle [j, j + 1]. Quant à la matrice diagonale η elle se définie comme nous l'avons évoqué précédemment, c'est-à-dire en sommant les lignes de $\hat{I}$. Avec ces notations et pour une série de taux universels $R = (r_0, r_1, \ldots, r_{T-1})'$ le résultat des calculs pour P_{ind} s'écrit en fait :

$$P_{ind} = \eta^{-1} \hat{I} R \qquad (8)$$

On retrouve ici une formule très familière…

2.3.3. La définition de l'indice

On peut maintenant donner une définition rigoureuse de l'indice, sans que celle-ci ne repose sur les hypothèses d'une régression. Elle aborde directement la question sous l'angle informationnel. Les modèles classiques (BMN et CS) sont des cas particuliers de cette définition générale.

Définition

Pour un échantillon de ventes répétées :

A partir des séries de taux $\{r_k(s)\}$ de chaque bien et des distributions informationnelles associées $\{inf_k(s)\}$ on calcule le vecteur réel des taux moyens, $P_{réel}$.
D'autre part, à chaque série de taux universels $R = (r_0, r_1, \ldots, r_{T-1})'$ est associé un vecteur indiciel des taux moyens P_{ind}.

Le vecteur des taux de croissance monopériodiques de l'indice est celui qui reproduit le vecteur réel des taux moyens. C'est-à-dire celui qui vérifie :

$$P_{ind} = P_{réel} \quad \Leftrightarrow \quad \eta^{-1}\, \hat{I}\, R = P_{réel} \quad \Leftrightarrow \hat{I}\, R = \eta\, P_{réel} \qquad (9)$$

La notion de moyenne informationnelle des taux moyens apparaît donc comme une notion centrale pour définir l'indice. L'idée de cette définition consiste à trouver la bonne série universelle qui rend compte exactement du vecteur des taux moyens réels, $P_{réel}$.

2.4. Exemples

Nous illustrerons ici la manière dont il est possible d'intégrer les ventes multiples et les valeurs d'expertise à l'indice. Puis, un exemple de calcul de matrice informationnelle sera exposé.

2.4.1. Un exemple de ventes multiples avec évaluation intermédiaire

Supposons qu'un bien k ait été acheté à la date t = 0 pour 100 000, revendu à la date t = 10 pour 150 000, évalué à la date t = 18 à 130 000, et enfin à nouveau revendu à la date t = 30 pour 180 000.

Pour définir la série des $\{r_k(s)\}_{s=0,\ldots,29}$ nous calculons sur chacun des intervalles [0,10], [10,18] et [18,30] le taux moyen réalisé, avec les prix ou avec la valeur d'expertise. On obtient 4,1% dans le premier cas[3], - 1,6% dans le second[4] et 2,7% pour le dernier intervalle[5]. On a ainsi :

$r_k(0) = \ldots = r_k(9) = 4{,}1\%$ $\quad r_k(10) = \ldots = r_k(17) = -1{,}6\%$ $\quad r_k(18) = \ldots = r_k(29) = 2{,}7\%$

Pour définir les contributions informationnelles $\{inf_{k'}(s)\}_{s=0,\ldots,29}$ on fera les hypothèses suivantes :

- Les intervalles élémentaires ayant une de leurs extrémités correspondant à une date de transaction, auront un niveau d'information fixé à 1 :

$$inf_k(0) = inf_k(9) = inf_k(10) = inf_k(29) = 1$$

- Une valeur d'expertise étant moins informative que l'observation d'un prix, les intervalles contigus à une date d'expertise auront un niveau d'information[6] de 0,9 :

$$inf_k(17) = inf_k(18) = 0{,}9$$

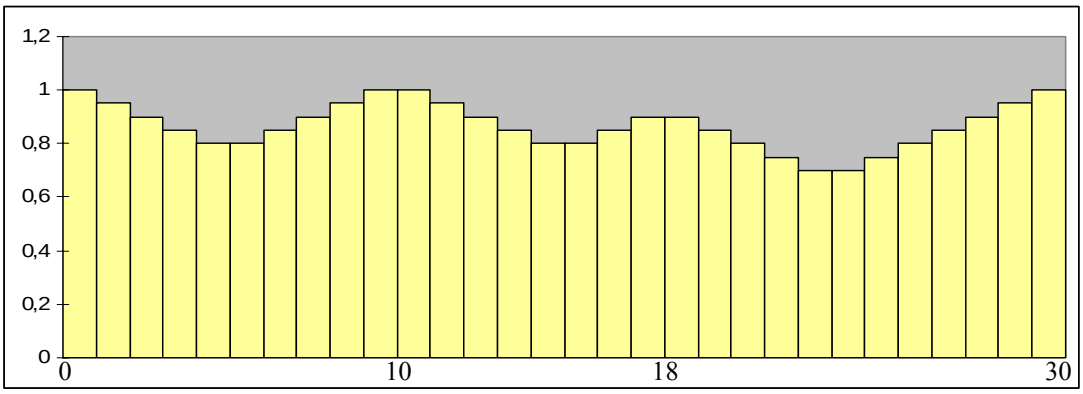
- A partir de ces valeurs, les niveaux d'information décroîtront de 0.05 en 0.05, au fur et à mesure que l'on s'éloignera des dates des transactions ou de l'expertise.

La figure 1 représente sous forme d'histogramme la série des $\{inf_k(s)\}$. Grâce à la généralisation informationnelle nous pouvons donc inclure dans un indice de type RSI les ventes multiples et les valeurs d'expertise, en pondérant leur influence

[3] $\text{Ln}(150000/100000) / 10 \approx 4{,}1\%$
[4] $\text{Ln}(130000/150000) / 8 \approx -1{,}6\%$
[5] $\text{Ln}(180000/130000) / 12 \approx 2{,}7\%$
[6] Le choix de ce 0,9 peut se justifier en rappelant qu'une évaluation est en général considérée comme correcte quand l'écart entre le vrai prix et la valeur d'expertise est inférieur à 10%.

respective. On pourra se référer à Hordijk, De Kroon, Theebe (2004) pour avoir un exemple d'indice hybride intégrant les ventes multiples et les évaluations fournies par les experts.

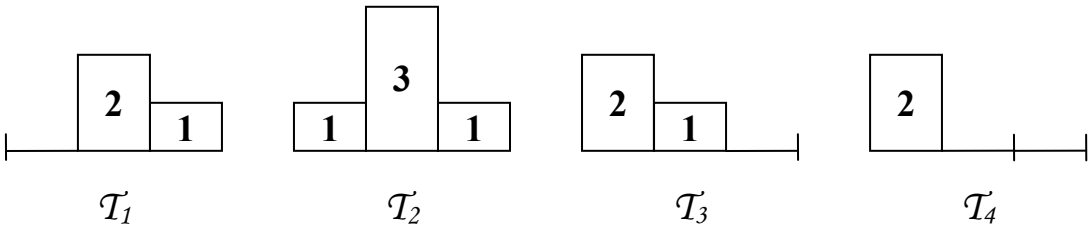
Figure 1 : Contributions informationnelles du bien k



2.4.2. Un exemple de matrice informationnelle

Afin d'appréhender plus concrètement la nouvelle structure de la distribution informationnelle nous supposerons dans ce deuxième exemple que le temps est limité à quatre dates (t = 0, 1, 2, 3). L'échantillon d'estimation ne comportera que quatre transactions $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ et $\mathcal{T}_4$ avec $\mathcal{T}_1 \in C(1,3)$, $\mathcal{T}_2 \in C(0,3)$, $\mathcal{T}_3 \in C(0,2)$ et $\mathcal{T}_4 \in C(0,1)$. Les diagrammes ci-dessous représentent les quantités d'information $inf_{k'}(s)$ fournies par chacune d'elles.

Figure 2 : Quantités d'information fournies par $\mathcal{T}_1$, $\mathcal{T}_2$, $\mathcal{T}_3$ et $\mathcal{T}_4$



Le tableau des $L_{i,j}$ comporte dans chaque case autant de valeurs qu'il y a d'unités de temps entre i et j pour rendre compte des séries $L_{i,j}(s)$.

Tableau 1 : Tableau des $L_{i,j}(s)$

$L_{0,1}(0) = 2$	$L_{0,2}(0) = 2$ $L_{0,2}(1) = 1$	$L_{0,3}(0) = 1$ $L_{0,3}(1) = 3$ $L_{0,3}(2) = 1$
	$L_{1,2}(1) = 0$	$L_{1,3}(1) = 2$ $L_{1,3}(2) = 1$
		$L_{2,3}(2) = 0$

Le passage de la distribution des $L_{i,j}$ à la matrice Î se fait d'une manière très semblable à ce que l'on a déjà rencontré précédemment. Pour chaque intervalle de temps unitaire [t, t+1] on sélectionne la population des ventes répétées pertinentes. On additionne ensuite les valeurs $L_{i,j}(0)$ entre elles, les valeurs $L_{i,j}(1)$ entre elles,…, jusqu'aux $L_{i,j}(T-1)$. Ces sommes fournissent les T éléments de la $t^{ème}$ ligne de Î. Dans le tableau 1 on a ainsi encadré en traits épais la population des ventes répétées pertinentes pour [1,2] et en sommant ces quantités comme indiqué on établit que la deuxième ligne de Î est : (3 ; 6 ; 2). Plus globalement la matrice Î est :

$$\hat{I} = \begin{pmatrix} 5 & 4 & 1 \\ 3 & 6 & 2 \\ 1 & 5 & 2 \end{pmatrix}$$

Cet exemple permet de remarquer immédiatement que certaines des propriétés des matrices Î, énoncées dans les propositions du chapitre 3, paragraphe 2.1, ne sont plus vraies dans le cadre généralisé. Ainsi, Î n'est plus symétrique et ses valeurs ne décroissent plus nécessairement en partant des éléments diagonaux comme on peut le constater avec la troisième ligne.

2.5. Comment définir l'information ?

L'impact de la quantification informationnelle sur les valeurs indicielles est examiné dans le premier paragraphe. Des exemples de définitions alternatives de l'information seront ensuite exposés.

2.5.1. Les enjeux

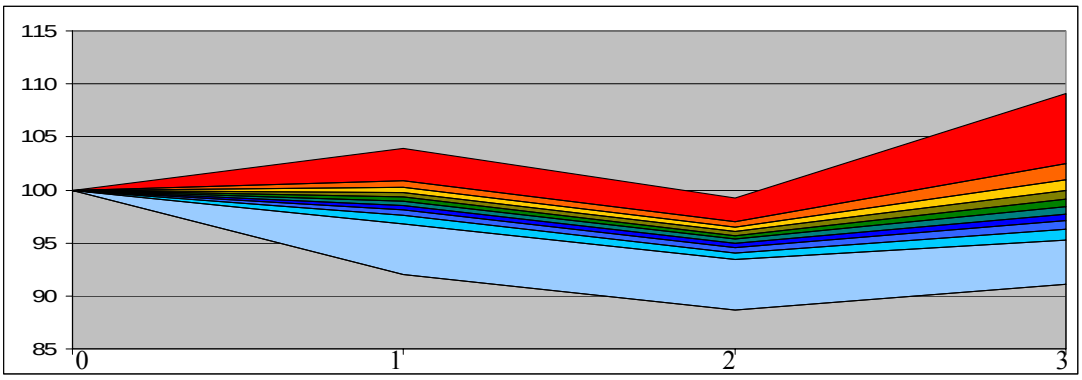
Au tout début de cette thèse, les difficultés théoriques auxquelles il faut faire face lorsque l'on cherche à définir la notion de prix de marché dans le cas des marchés hétérogènes et illiquides ont été discutées. Il a ainsi été mis en avant que ce concept était difficile à fonder rigoureusement, mais incontournable empiriquement. Le principal problème, à savoir la non fongibilité, pourrait être résumé d'une manière un peu caricaturale, par la question suivante : Combien y a-t-il d'immobilier dans une maison ou un appartement ? Ou, en d'autres termes, quel est le degré de représentativité d'un bien particulier vis-à-vis de cette matière première abstraite que l'on appelle « l'immobilier » et que l'indice cherche à retranscrire ? Répondre à cette question revient en fait à déterminer le niveau informationnel d'un bien ce qui, avec les notations introduites précédemment, équivaut à préciser les valeurs des séries $\{inf_k(s)\}$ (dans le cas d'un modèle de ventes répétées).

Si, du point de vue formel, mettre en évidence la structure informationnelle implicite au RSI constitue un progrès, toutes les difficultés ne sont pas pour autant résolues car dorénavant une question cruciale se pose : Comment faut-il définir

l'information ? Le modèle de Case, Shiller est-il le meilleur ou faut-il choisir d'autres pondérations informationnelles ? Cette interrogation devient maintenant le centre de la problématique et selon les réponses que l'on y apportera le prix de marché de l'immobilier, c'est-à-dire la courbe indicielle, variera.

Pour illustrer les effets de la définition de l'information sur un indice on réalise l'expérience suivante. On suppose que le temps est à nouveau limité à 4 dates (t = 0,1,2,3) et on constitue un échantillon de 13 ventes répétées pour lesquelles les dates d'achat et de revente sont connues, ainsi que les prix des transactions (cf. annexe 24). On suppose de plus que les $\{inf_k(s)\}$ sont constants sur chaque intervalle élémentaire des périodes de détention, comme dans le modèle Case-Shiller classique, mais au lieu de définir l'information apportée par une donnée avec la formule $(\Theta+(j-i))^{-1}$ on lui permet de varier aléatoirement[7] entre 0,2 et 1. Une fois ces niveaux fixés pour les 13 ventes répétées, l'indice est calculé. Si on génère ensuite, toujours aléatoirement, une deuxième distribution de l'information on obtient une deuxième série de valeurs indicielles, très probablement différentes de celles obtenues dans le premier cas. Quelle est alors l'ampleur des variations ? Pour obtenir la réponse nous simulons 30000 distributions informationnelles, tout en gardant les mêmes prix pour les 13 transactions. A chaque date les valeurs des indices sont ordonnées par déciles et les résultats que l'on obtient sont présentés dans la figure 3.

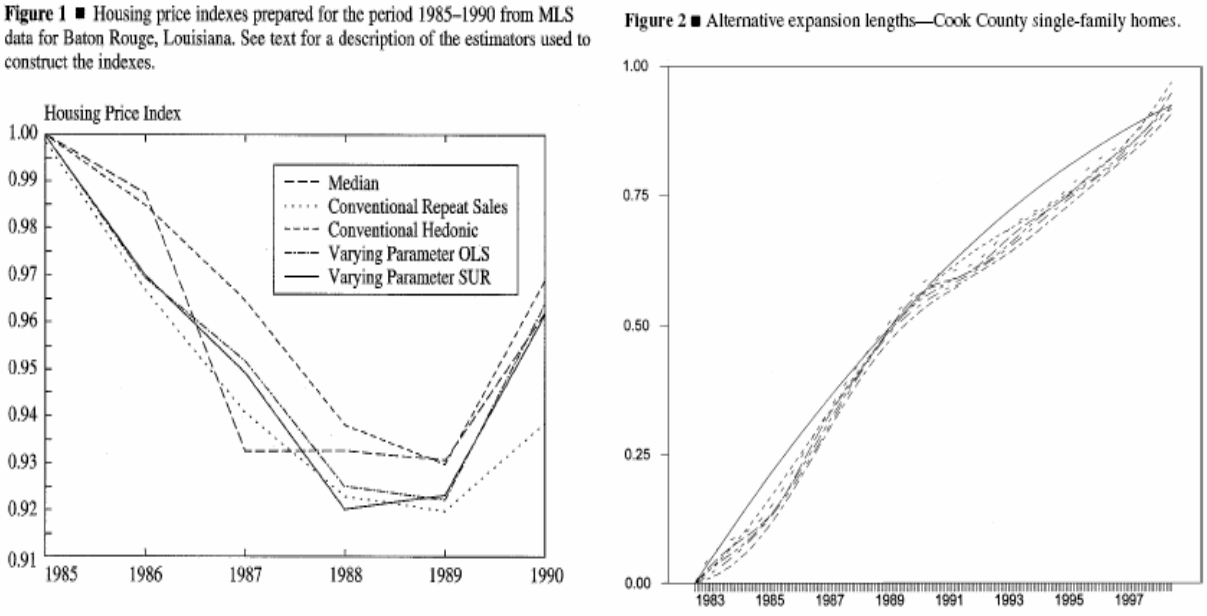
Figure 3 : Déciles indiciels quand l'information varie entre 0,2 et 1



[7] $inf_k(s) = inf_k$ est fixé aléatoirement pour chaque bien k grâce à la formule 0,2 + 0,8U où U suit une loi aléatoire uniforme sur [0,1]

Une courbe indicielle est en fait une courbe qui serpente dans cette bande de valeurs potentielles. A une date donnée, 10% des valeurs seront dans la zone bleue et 10% dans la zone rouge, ces deux portions constituent les déciles extrêmes. A la date t = 3 les résultats varient entre 91 et 109, la définition de l'information impacte donc très significativement les résultats du calcul. En parcourant les articles de la littérature on retrouve fréquemment ce genre d'illustration. Le graphique de gauche de la figure 4 présente les courbes que l'on obtient en calculant différents indices hybrides (Knight, Dombrow, Sirmans (1995)). Selon la méthodologie employée, ou en d'autres termes selon la définition de l'information choisie implicitement, les résultats semblent varier dans une bande indicielle. Le graphique de droite, issu de McMillen, Dombrow (2001), fournit également le même genre d'intuition[8]. Il en est de même pour la figure 5, extraite de Cannaday, Munneke, Yang (2005), où l'information est ici implicitement une fonction de l'age de la propriété considérée.

Figure 4 : Deux exemples de ruban indiciel



Knight, Dombrow, Sirmans (1995) McMillen, Dombrow (2001)

[8] On ne cherchera pas ici à établir les structures d'information implicites aux modèles associés à ces articles car elle sont probablement très complexes. On se bornera simplement à vérifier visuellement l'intuition d'une fluctuation dans une bande de valeurs indicielles.

Figure 5 : Exemples de ruban indiciel : l'information est fonction de l'age

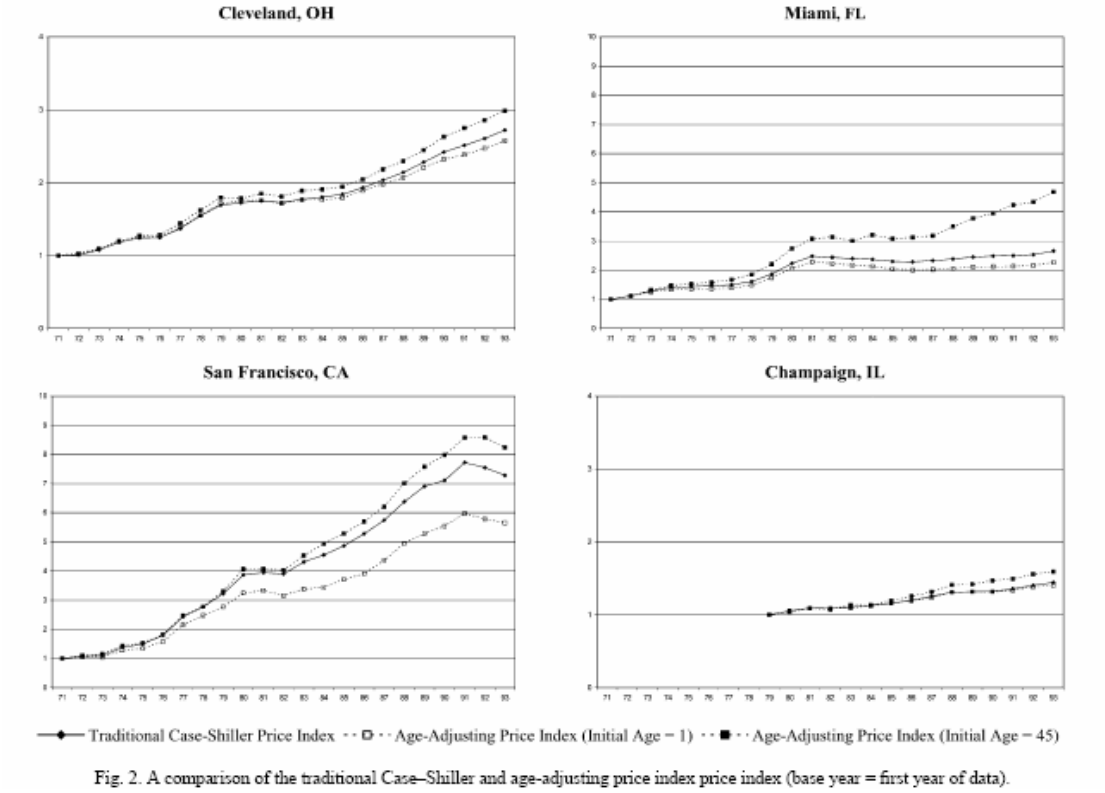


Fig. 2. A comparison of the traditional Case–Shiller and age-adjusting price index price index (base year = first year of data).

Cannaday, Munneke, Yang (2005)

2.5.2. Les définitions alternatives de l'information utilisées dans la littérature

Depuis Case, Shiller (1987) de nombreuses variantes du modèle des ventes répétées ont été proposées et étudiées. L'idée sous-jacente et implicite, commune à tous ces articles, consiste en fait à choisir une autre définition de l'information. Une fois que la méthode de quantification de cette grandeur est déterminée la structure de l'indice reste souvent la même (la relation $\hat{I}R = \eta P$ est en général toujours valide). La réflexion développée dans cette thèse permet donc d'unifier des méthodes et des résultats qui auraient pu apparaître disparates et indépendants sans cela.

L'article de Peng (2000), dans la suite des travaux de Shiller (1991), propose ainsi plusieurs types d'indices arithmétiques : équipondéré, pondéré par les prix,

pondéré par les valeurs. Ces modèles, très intéressants sur le plan de la gestion des portefeuilles immobiliers, ne consistent en fait, sur un plan plus fondamental, qu'en une simple redéfinition des $L_{i,j}$.

Dans le même ordre d'idée, l'article de Dreiman, Pennington-Cross (2004) remet en question et approfondit la modélisation classique du résidu ε (bruit blanc et marche aléatoire gaussienne). Les auteurs proposent d'introduire un terme quadratique dans sa variance et d'estimer l'indice sur cette base. Dans le modèle traditionnel de Case et Shiller la quantité d'information fournie par un bien, acheté à t_i et revendu à t_j, se calculait par $(\Theta + (j - i))^{-1}$. La méthode proposée par Dreiman et Pennington-Cross revient à calculer cette quantité par $(\Theta + \psi (j - i) + (j - i)^2)^{-1}$, tout en gardant le reste de la structure du RSI inchangé. On pourra également consulter l'article de Hill, Sirmans, Knight (1999) où la modélisation de la variance du terme d'erreur fait intervenir un processus de type AR(1).

3. Une géométrie de l'information

Dans ce paragraphe nous nous placerons à nouveau dans un contexte classique de type Case-Shiller, tel que présenté dans le paragraphe 6 du chapitre 1, et on explorera les propriétés de la distribution de l'information à l'aide de concepts géométriques.

3.1. La représentation spatiale de Î

A chaque matrice Î est associé un ensemble de T vecteurs, cette section étudie leurs propriétés.

3.1.1. Notations

L'étude de la matrice informationnelle Î peut fournir des éléments intéressants pour l'analyse si on l'aborde sous l'angle géométrique. Cette matrice peut s'écrire

(v_0, v_1, … , v_{T-1}), où les $\{v_i\}$ sont des vecteurs colonnes de dimension T. Un vecteur v_i est associé à l'ensemble des couples pertinents pour [i, i+1], la $j^{ème}$ composante de v_i donnant la quantité d'information fournie par cet ensemble de ventes répétées pour l'intervalle [j-1, j]. La $i^{ème}$ coordonnée de v_i est bien sûr la plus grande, mais l'information concernant les autres intervalles temporels n'est pas du tout insignifiante. Cette famille de vecteurs peut se représenter dans un espace Euclidien de dimension T, où les axes orthogonaux mesurent les contributions informationnelles pour les intervalles de temps élémentaires.

Les coordonnées des v_i étant toutes positives, ces vecteurs appartiennent tous au quadrant :

$$Q^+ = \{(x_1, \ldots, x_T) \; ; \; x_i \geq 0\}.$$

A l'intérieur de Q^+ plusieurs aires peuvent être délimitées[9] :

$$\mathcal{P}_i = Q^+ \cap \{ (x_1, \ldots, x_T) / \; x_i \geq x_1 \; ; \ldots ; x_i \geq x_T \} \qquad i = 1,\ldots,T$$

Un domaine $\mathcal{P}_i$ regroupe l'ensemble des vecteurs de Q^+ pour lesquels la $i^{ème}$ coordonnée est dominante. On a bien sûr :

$$Q^+ = \cup_{i=0,\ldots,T} \mathcal{P}_i$$

L'intersection de ces zones définit, quant à elle, une demi-droite :

$$d^+ = \cap_{i=0,\ldots,T} \mathcal{P}_i = \{ (x_1, \ldots, x_T) \; ; \; x_1 = x_2 = \ldots = x_T \}$$

Chaque vecteur v_i se situe dans sa zone $\mathcal{P}_i$ correspondante. La figure 6 présente une illustration de ces concepts en dimension T = 2.

[9] Ces $\mathcal{P}_i$ n'ont rien à voir avec les notations employées dans le paragraphe 4 du chapitre 3 pour désigner les deux populations dans la modélisation du même nom.

3.1.2. La contribution informationnelle moyenne du vecteur v_i

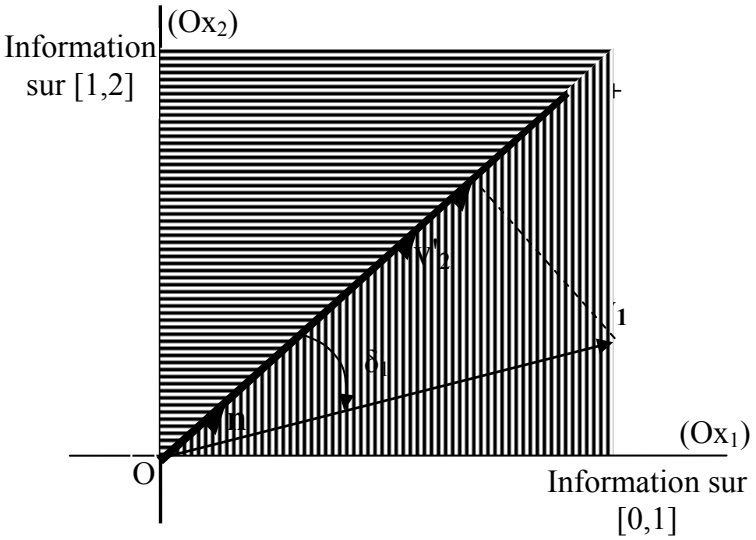
La valeur moyenne des coordonnées de $v_i = (v_i(1),\ldots,v_i(T))$ est $E(v_i) = (v_i(1) +\ldots+v_i(T))/T$, le moment d'ordre 2 s'écrit $E(v^2_i) = (v^2_i(1) + \ldots + v^2_i(T)) / T$ et la variance est $V(v_i) = E(v^2_i) - [E(v_i)]^2$. Pour le produit scalaire euclidien basique on a donc $\|v_i\|^2 = T\,E(v^2_i)$.

Un vecteur v_i se situe toujours entre son axe associé (Ox_i) et la demi-droite d^+, on notera par δ_i l'angle géométrique entre d^+ et v_i dans le plan engendré par (Ox_i) et d^+. On notera de plus par v'_i la projection orthogonale de v_i sur d^+ et par n le vecteur normalisé[10] de d^+. Si l'on remarque que $v'_i \,.\, n = v_i \,.\, n$, la norme de v'_i s'écrit alors :

$$\|v'_i\| = (\, v_i(1) + \ldots + v_i(T)\,) * T^{-½} = T^{½}\,E(v_i) \qquad (10)$$

La longueur de v'_i donne donc le niveau moyen de contribution informationnelle fournie par les ventes répétées pertinentes pour [i,i+1], pour tous les intervalles unitaires de [0,T] (au facteur $T^{½}$ près).

Figure 6 : Représentation spatiale des vecteurs informationnels



[10] Ses coordonnées sont toutes égales à $T^{-½}$

3.1.3. Mesure de la diffusion informationnelle du vecteur v_i

En utilisant Pythagore, la distance entre v_i et d^+, c'est-à-dire la troisième longueur du triangle (O,v_i,v'_i), peut se calculer par :

$$\| v_i \|^2 - \| v'_i \|^2 = T\, E(v^2_i) - T\, [E(v_i)]^2 = T\, V(v_i) \qquad (11)$$

Plus v_i est éloigné de d^+, plus la distribution des $\{v_i(j)\}$ est dispersée[11]. Si l'on souhaite comparer la dispersion de v_k avec la dispersion d'un autre vecteur $v_{k'}$, il est nécessaire de normaliser la mesure ; on utilisera donc plutôt l'angle δ_i que $V(v_i)$. Cet angle peut se calculer grâce à la formule trigonométrique classique :

$$tg\,(\delta_i) = \sigma(v_i) / E(v_i) = \left[\, E(v^2_i) / [E(v_i)]^2 - 1 \,\right]^{1/2} \qquad (12)$$

Il est intéressant de connaître pour un niveau fixé de $v_i(i) = I^{[i-1,i]}$, l'étendue des valeurs $[\delta_{inf}\ ,\ \delta_{sup}]$ que peut parcourir l'angle δ_i. On démontre dans l'annexe 25 que cet angle δ_i varie entre $\delta_{inf} = 0$ quand v_i est sur d^+, et $\delta_{sup} = tg^{-1}(\,(T-1)^{1/2}\,)$ quand v_i est sur[12] (Ox_i).

Dans la première situation l'ensemble de l'information pertinente pour [i-1,i], $v_i(i) = I^{[i-1,i]}$, est utile pour tous les intervalles. En d'autres termes, tous les achats ont été réalisés à t = 0 et toutes les reventes à t = T. Tandis que dans la seconde, $v_i(i)$ fournit de l'information uniquement pour l'intervalle [i-1,i] : tous les achats ont eu lieu à t = i – 1 et les toutes les reventes à t = i. On peut donc interpréter δ_i comme une mesure de la diffusion de l'information et aussi indirectement comme une mesure de la longueur de la période de détention. Dans la suite de ce paragraphe le terme "diffusante" sera utilisé pour désigner les ventes répétées fournissant de l'information pour un grand nombre d'intervalles (période de détention longue), le terme

[11] La dispersion étant mesurée par $V(v_i)$

[12] Le maximum est de 45° seulement en dimension 2, pour les dimensions supérieures cette valeur peut être dépassée.

"spécialisée" se référera par contre aux ventes associées à un faible nombre d'intervalles (période de détention courte).

3.1.4. Modification marginale de l'information fournie par v_i

On peut approfondir l'intuition sur la mesure δ_i en examinant la situation suivante. A partir d'un vecteur $v_i = (v_i(1), \ldots, v_i(T))$ donné, on fait varier la $j^{ème}$ composante ($j \neq i$) de $v_i(j)$ à $v_i(j) + \varepsilon$. L'augmentation marginale de $v_i(j)$ signifie qu'une part plus grande de $I^{[i-1,i]}$ devient pertinente pour le $j^{ème}$ intervalle. En dérivant la fonction cos (δ_i), et en introduisant la notation $diff(v_i) = E(v_i)\,[1 + tg^2(\delta_i)]$, on démontre dans l'annexe 26 qu'une augmentation marginale de $v_i(j)$ agit différemment sur le degré de diffusion informationnelle selon que $v_i(j)$ est au-dessus ou au-dessous du seuil $diff(v_i)$:

- si $v_i(j)$ est inférieure à $diff(v_i)$ la diffusion augmente
- inversement si $v_i(j) > diff(v_i)$, c'est-à-dire si $v_i(j)$ est proche de $v_i(i)$ et donc assez éloigné de la valeur moyenne $E(v_i)$, l'information deviendra plus spécialisée.

Par exemple, si le vecteur v_3 est (2 ; 4 ; 5 ; 3 ; 1) on a $diff(v_3) \approx 3.67$. Le vecteur w_3 = (2 ; 4.1 ; 5 ; 3 ; 1) est alors informationnellement plus spécialisé que v_3, tandis que le vecteur x_3 = (2 ; 4 ; 5 ; 3.1 ; 1) est informationnellement plus diffusant.

La mesure δ_i ne compare pas en fait le poids informationnel du seul intervalle [i-1,i] par rapport aux poids des autres intervalles. Elle compare plutôt le poids informationnel des intervalles proches de [i-1,i], c'est-à-dire ceux associés à la condition $v_i(j) > diff(v_i)$, et le poids des intervalles plus éloignés.

3.2. Les volumes informationnels

Après avoir étudié les caractéristiques des vecteurs v_i isolément, on les considère dans leur ensemble. On aboutira ici à la notion de volume informationnel.

3.2.1. Les volumes élémentaires $\mathcal{R}_1$ et $\mathcal{R}_2$.

Les vecteurs colonnes de la matrice carrée Î peuvent donc se représenter dans un espace de dimension T. Comme on l'a déjà mentionné, le vecteur v_t est principalement associé au $t^{ème}$ intervalle de temps mais il est aussi secondairement utile pour les autres intervalles[13]. Ces T vecteurs peuvent être vus comme les côtés d'un parallélépipède $\mathcal{R}$ de dimension T (cf. parallélogramme de la figure 7, pour T = 2). Les coordonnées de son sommet P sont fournies par les sommes sur les lignes de la matrice Î. On obtient donc le nombre ajusté[14] de couples pertinents pour chaque intervalle de temps $n^t\, G(\zeta^t)$. Et comme $\tau^t = (n^t\, G(\zeta^t)) / I^t$, on a immédiatement la moyenne des périodes de détention τ^t en divisant la $t^{ème}$ coordonnée de P par la $t^{ème}$ coordonnée de v_t.

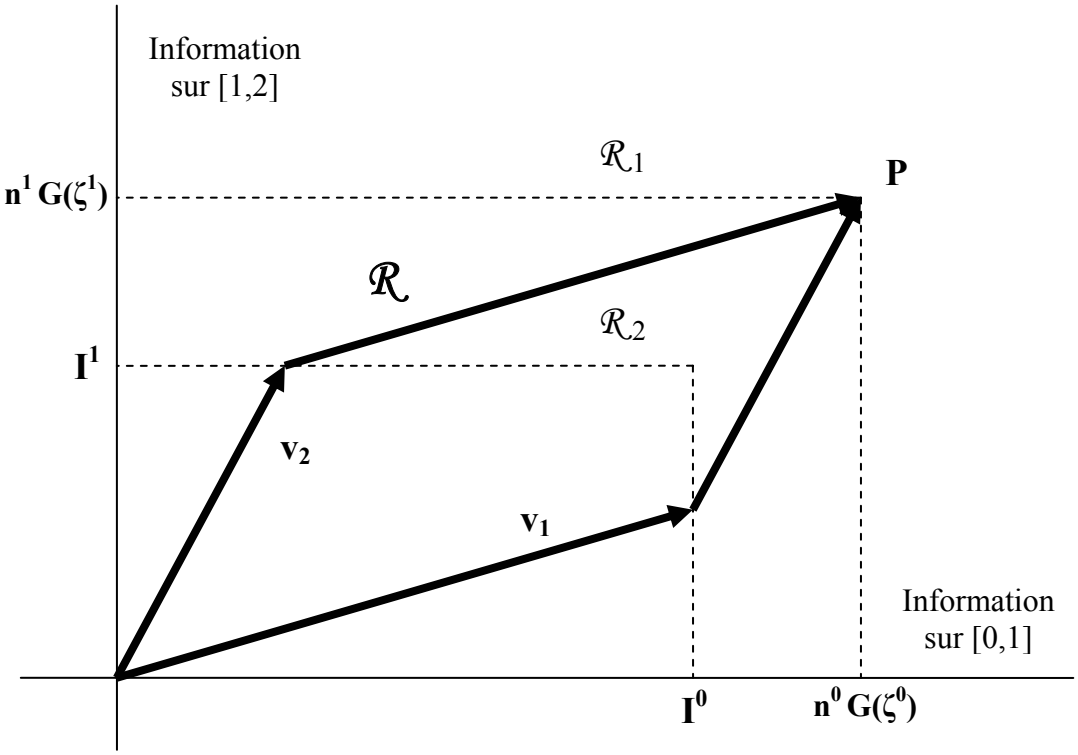
Les quantités $n^t\, G(\zeta^t)$ et I^t pointées sur chaque axe (Ox_t) peuvent aussi être vues comme les longueurs des côtés de deux parallélépipèdes rectangles $\mathcal{R}_1$ et $\mathcal{R}_2$, avec $\mathcal{R}_2 \subset \mathcal{R}_1$. Les volumes sont fournis par les déterminants des matrices dont les vecteurs colonnes correspondent aux côtés de ces parallélépipèdes. Pour $\mathcal{R}_1$ la matrice associée est donc simplement η et pour $\mathcal{R}_2$ il s'agit de la matrice diagonale construite à partir des nombres $(I^0, I^1, \ldots, I^{T-1})$. L'inclusion $\mathcal{R}_2 \subset \mathcal{R}_1$ amène alors :

$$\prod_{i=0,\ldots,T-1} I^i \leq \prod_{i=0,\ldots,T-1} n^i\, G(\zeta^i) = \det(\eta) \qquad (13)$$

[13] Sa t^{eme} coordonnée est dominante mais les autres ne sont pas pour autant nulles, ou même négligeables

[14] Ajustement par $G(\zeta^t)$.

Figure 7 : Volumes informationnels



Pour les parallélépipèdes rectangles $\mathcal{R}_1$ et $\mathcal{R}_2$, l'utilisation d'un déterminant pour obtenir les volumes peut sembler excessivement sophistiqué car le résultat est évident. Cette technique prend par contre tout son sens quand on veut calculer le volume du parallélépipède $\mathcal{R}$ construit sur les vecteurs (v_1 , … , v_T). La matrice associée étant Î, ce volume non trivial s'obtient immédiatement en calculant det(Î).

3.2.2. L'interprétation du volume informationnel $\mathcal{R}$

Avant d'aller plus loin dans l'étude des relations entre ces différents volumes il n'est pas inutile d'illustrer par un exemple la signification de la mesure det(Î). Les tableaux suivants présentent deux distributions de $\{L_{i,j}\}$, pour T = 2.

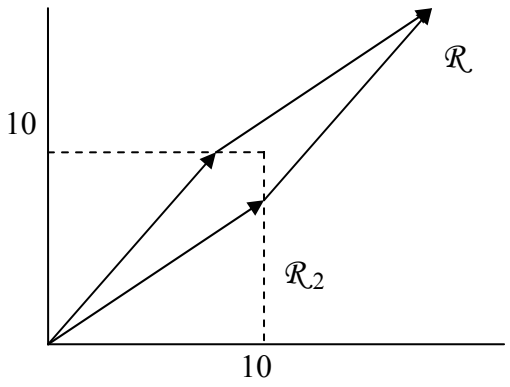
1	9
	1

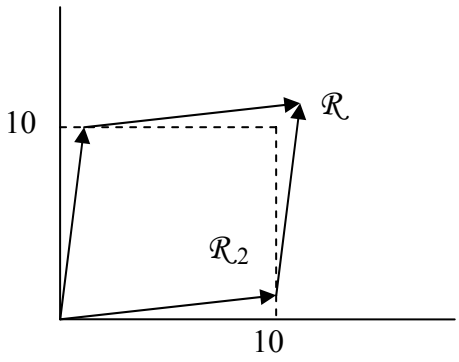
9	1
	9

Les matrices informationnelles sont :

$$\begin{pmatrix} 10 & 9 \\ 9 & 10 \end{pmatrix} \qquad \begin{pmatrix} 10 & 1 \\ 1 & 10 \end{pmatrix}$$

Et géométriquement les parallélogrammes associés donnent les figures suivantes :





Ces deux cas présentent les mêmes niveaux d'information I^0 et I^1 pour les intervalles [0,1] et [1,2], mais les situations sont toutefois très différentes. Le premier parallélogramme est étiré et fin tandis que le second est presque de forme carrée ; si l'on calcule les déterminants associés on obtient respectivement 19 et 99. Un parallélogramme fin signifie que les vecteurs v_i sont proches de la demie droite d^+ et que les angles δ_i sont petits, en d'autres termes les périodes de détention sont longues et l'information est fortement diffusée. Inversement une forme quasi-rectangulaire signifie que les angles δ_i sont grands et que les périodes de détention sont courtes, dans ce cas l'information est fortement spécialisée sur un petit nombre d'intervalles. Le premier cas engendre donc un petit déterminant, correspondant à une situation où

le poids de la distribution des $\{L_{i,j}\}$ est concentré sur le coin supérieur droit, tandis que le second cas produit un déterminant important témoignant d'une distribution des $\{L_{i,j}\}$ pour laquelle la masse des données est plutôt proche de la diagonale :

Synthétiquement, det($\hat{I}$) est donc un indicateur mesurant la diffusion de l'information ou, de manière équivalente, la longueur des périodes de détention .

3.2.3. Maximum et minimum atteints par le volume $\mathcal{V}ol(\mathcal{R})$

Cet indicateur de la diffusion informationnelle qu'est le volume de $\mathcal{R}$, peut s'étudier en utilisant l'inégalité de Hadamard pour les matrices symétriques définies positives. On obtient alors :

$$0 \leq \det(\hat{I}) \leq \prod_{i=0,\ldots,T-1} I^{i}$$

Et en y ajoutant les volumes $\mathcal{R}_1$ et $\mathcal{R}_2$ on arrive à :

$$0 \leq \mathcal{V}ol(\mathcal{R}) \leq \mathcal{V}ol(\mathcal{R}_2) \leq \mathcal{V}ol(\mathcal{R}_1) \qquad (14)$$

La borne supérieure $\mathcal{V}ol_{max} = \prod_{i=0,\ldots,T-1} I^{i} = \mathcal{V}ol(\mathcal{R}_2)$ est atteinte dans l'inégalité de Hadamard si et seulement si la matrice $\hat{I}$ est diagonale, c'est-à-dire pour $\hat{I} = \hat{I}_{max} = \text{diag}(I^0, I^1, \ldots, I^{T-1})$. En termes géométriques cette situation signifie que les vecteurs v_i sont sur leur axe respectif (Ox_i) et que l'on a $\mathcal{R} = \mathcal{R}_2$.

La borne inférieure dans la formule (14) est 0, mais il est probablement possible d'améliorer cette minoration en approfondissant l'étude de la matrice Î. Pour chaque élément I^i du vecteur (I^0, I^1, … , I^{T-1}) nous définissons les quantités FLIV(I^i) et FRIV(I^i) par :

FLIV(I^i) = I^k si (k < i) et (pour tous les j = k+1, … , i – 1 on a : $I^j > I^i$)

FRIV(I^i) = I^k si (k > i) et (pour tous les j = i+1, … , k – 1 on a : $I^j > I^i$)

(FLIV : “first left inferior value” FRIV : “first right inferior value”)

Lorsque ce I^k n'existe pas, ce qui se produit par exemple pour I^0 et I^{T-1} quand on veut calculer FLIV(I^0) et FRIV(I^{T-1}), nous fixons les valeurs FLIV et FRIV à 0. A partir de ces deux grandeurs on définit ensuite FMIV(I^i) par :

FMIV(I^i) = Max (FLIV(I^i) , FRIV(I^i)) (“first minimum inferior value”)

Nous admettrons alors le résultat suivant sans démonstration :

Conjecture

La borne inférieure pour det(Î) est : $\mathcal{V}ol_{min} = \prod_{i=0,\dots,T-1} (I^i - FMIV(I^i))$ (15)

Elle est atteinte si et seulement si Î correspond à la matrice symétrique $\hat{I}_{min}$ définie par les relations :

$\hat{I}_{min}(p,p) = I^{p-1}$ et $\hat{I}_{min}(p,q) = Min (\hat{I}_{min}(p,q-1) ; \hat{I}_{min}(p+1,q))$ pour $1 \le p < q < T$

La matrice I_{min} peut s'interpréter comme la matrice de diffusion maximale (ou matrice de spécialisation minimale). En effet, à partir d'un élément diagonal I^{p-1}

(quantité d'information pertinente pour [p-1, p]), si l'on se décale vers le haut ou vers la droite, la proposition 1 (chapitre 3, paragraphe 2.1) affirme que la valeur rencontrée doit être inférieure à I^{p-1}. Plus généralement il en est de même si l'on part d'une cellule strictement au-dessus de la diagonale. En termes concrets cette propriété signifie qu'une partie de l'information pertinente pour un intervalle est toujours pertinente si l'intervalle est étendu par la droite ou par la gauche, mais que cette portion est nécessairement inférieure à l'information initiale. L'idée sous-tendant la matrice I_{min} est alors de satisfaire à cette règle tout en gardant les différentes quantités d'information à leur niveau maximal autorisé ; l'information est ici fortement diffusée. Il faut toutefois remarquer que la matrice I_{min} doit aussi satisfaire (et c'est effectivement le cas) à l'inégalité découlant de la proposition 2 (chapitre 3, paragraphe 2.1). Enfin, puisque la matrice I_{min} a été qualifiée de matrice à diffusion maximale, nous parlerons pour I_{max} de matrice à diffusion minimale (ou à spécialisation maximale)

Si $\mathcal{Vol}(\mathcal{R})$ est proche de $\mathcal{Vol}_{min}$, le parallélépipède $\mathcal{R}$ est étiré et les périodes de détentions sont longues. Inversement plus $\mathcal{Vol}(\mathcal{R})$ est proche de $\mathcal{Vol}_{max} = \mathcal{Vol}(\mathcal{R}_2)$, plus le volume $\mathcal{R}$ est rectangle et plus les périodes de détention sont courtes. On obtient ainsi un nouvel indicateur global pour la longueur de détention, à savoir le ratio $\mathcal{Vol}(\mathcal{R}) / \mathcal{Vol}(\mathcal{R}_2)$. Si l'on utilise les propriétés classiques du déterminant et en particulier sa multilinéarité, on peut s'apercevoir que ce ratio n'est rien d'autre que det(J), où J est la matrice de dispersion (cf. chapitre 1, paragraphe 6.3.3). En utilisant la formule (15) qui donne la valeur de $\mathcal{Vol}_{min}$, les inégalités d'encadrement de $\mathcal{Vol}(\mathcal{R})$ se réécrivent :

$$\mathcal{Vol}_{min} \leq \mathcal{Vol}(\mathcal{R}) \leq \mathcal{Vol}_{max} \quad \Leftrightarrow \quad \prod_{i=0,\ldots,T-1}\left(1 - FMIV(I^i)/I^i\right) \leq \det(J) \leq 1 \quad (16)$$

L'indicateur det(J) varie donc entre $\prod_{i=0,\ldots,T-1}(1 - FMIV(I^i)/I^i)$ et 1.

3.2.4. Un exemple numérique

A titre d'illustration si la matrice Î est :

$$\hat{\mathbf{I}} = \begin{pmatrix} 7 & 6 & 2 & 2 \\ 6 & 9 & 3 & 2 \\ 2 & 3 & 5 & 4 \\ 2 & 2 & 4 & 12 \end{pmatrix}$$

On a alors :

$$\hat{\mathbf{I}}_{\mathbf{max}} = \begin{pmatrix} 7 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 12 \end{pmatrix} \qquad \hat{\mathbf{I}}_{\mathbf{min}} = \begin{pmatrix} 7 & 7 & 5 & 5 \\ 7 & 9 & 5 & 5 \\ 5 & 5 & 5 & 5 \\ 5 & 5 & 5 & 12 \end{pmatrix}$$

Les valeurs inférieures associées au vecteur $(I^0, I^1, I^2, I^3) = (7 , 9 , 5 , 12)$ sont :

Tableau 2 : Valeurs des grandeurs FLIV, FRIV, FMIV

i	0	1	2	3
FLIV(I^i)	0	7	0	5
FRIV(I^i)	5	5	0	0
FMIV(I^i)	5	7	0	5

D'où : $\mathcal{V}ol_{\min} = (7 - 5) (9 - 7) (5 - 0) (12 - 5) = 140 \qquad \mathcal{V}ol_{\max} = 3780$

et : $\mathcal{V}ol(\mathcal{R}) = \det (\hat{I}) = 932 \qquad (140 < 932 < 3780)$

L'inégalité $\mathcal{V}ol_{\min} \leq \mathcal{V}ol(\mathcal{R}) \leq \mathcal{V}ol_{\max}$ est donc bien satisfaite et l'indicateur det(J) vaut 932/3780 ≈ 0,25.

4. Immobilier et théorie de l'arbitrage ? (Le cas des prêts hypothécaires)

Si aujourd'hui la financiarisation de l'immobilier semble se confirmer, il faut cependant prendre garde de ne pas transférer abusivement des méthodes financières classiques à ce champ d'application particulier, sans s'interroger préalablement sur la légitimité scientifique de ce transfert. Ce paragraphe étudie ainsi les conséquences d'un emploi abusif de la théorie de l'arbitrage à l'évaluation des options de remboursement anticipé, ou de défaut, incluses dans les prêts immobiliers hypothécaires. Les difficultés seront d'abord analysées théoriquement, en reprenant une démonstration classique d'arbitrage et en pointant les conditions requises par la théorie qui ne sont pas satisfaites[15] lorsqu'un des actifs sous-jacents est un bien immobilier. Beaucoup d'articles traitant de la modélisation des « mortgages[16] » considèrent en effet ces conditions comme automatiquement vérifiées et appliquent alors, très mécaniquement[17], les résultats de la théorie. Dans un deuxième temps si l'on souhaite malgré tout utiliser cette méthode, en étant bien conscient que le prix obtenu ne sera qu'approximatif, il faut pouvoir disposer au minimum d'une certaine intuition de l'erreur d'évaluation potentielle. En prenant l'exemple de l'option de vente dans le prêt viager hypothécaire on démontrera, malheureusement, que cette erreur peut être très conséquente. Ainsi, tant du point de vue théorique que du point

[15] Unicité du bien et non fongibilité, impossibilité de réaliser des positions shorts, incertitude sur le prix en l'absence de transaction effective, impossibilité de construire un portefeuille risque-neutre …
[16] Prêts immobiliers hypothécaires.
[17] Et sans doute même trop mécaniquement

de vue empirique, on vérifiera que la théorie de l'arbitrage ne semble pas pouvoir s'appliquer, au moins sous cette forme, à des sous-jacents immobiliers.

Ce paragraphe, un peu singulier par rapport au reste de la thèse, devra d'abord être compris comme une réflexion méthodologique ou comme une sorte de credo scientifique affirmant que les méthodologies d'étude doivent être fondamentalement adaptées à leur objet, les transferts trop rapides de techniques étant la source d'erreurs non négligeables.
Une des principales difficultés de l'arbitrage appliqué à l'immobilier est la méconnaissance du prix de marché (conceptuellement et empiriquement). Or, si l'actif primitif est mal connu, l'actif contingent le sera alors encore plus. Avant d'essayer d'évaluer des options, ou dans un avenir proche des produits dérivés sur indices immobiliers, il est donc nécessaire d'améliorer la mesure du prix ou du rendement immobilier. Dans cette optique, l'étude des indices s'impose d'elle-même. Ces instruments constituent la voie la plus naturelle pour financiariser et quantifier cet actif capricieux, illiquide et hétérogène qu'est l'immobilier. Ce n'est que dans un deuxième temps et à partir d'un socle défini rigoureusement, que l'on pourra ensuite tenter d'évaluer les produits dérivés. Economiquement parlant, il ne se s'agit pas là d'un enjeu mineur[18]. Ce paragraphe n'ira pas jusqu'à développer des méthodes de valorisation explicites, il cherchera simplement à mettre au jour les difficultés inhérentes à cet exercice quand l'actif immobilier est en jeu.

4.1. Introduction

For some years now, real estate and financial theory have become more and more interlinked. The approach of applying the powerful tools developed in finance to real estate seems promising; mortgage valuation methods developed by Kau, Keenan, Muller, Epperson (1992) is a major example. That article applied arbitrage theory

[18] Sur le marché de Londres en 2005, l'encours des dérivés immobiliers représentait 1 milliard d'euros

with two state variables (interest rate and a house process) to price a mortgage, which constitutes of a loan, a prepayment option and a default option. Some well-known arguments lead to a valuation PDE (partial differential equation) which is solved numerically.

The underlying assumptions of financial theory (for instance the absence of arbitrage opportunities) are clearly based on market reality. They are not purely formal or only useful to solve the mathematical problems that arise, they are leaning on the distinctive characteristics of the studied object. The purpose of this section is to re-examine the financial arbitrage models from the point of view of the specific features of real estate, in order to see if the requirements of arbitrage theory are entirely satisfied within this particular asset. This specific question can be compared to the case of real options theory, in which the utilisation of financial models was not initially straightforward.

The remainder of this paragraph is organized as follows. In the section 4.2, the optional model for mortgage is presented in a detailed manner. After a literature review, the demonstration leading to the fundamental valuation PDE is reviewed, trying to be as explicit as possible whenever a market assumption is necessary. Section 4.3 goes back to the demonstration and examines the relevance of assumptions pointed out in section 4.2. Does the riskless portfolio exist when dealing with real estate? Can real arbitrage activities exist and which assets could be chosen to build an actual arbitrage portfolio? These practical questions have direct consequences to the correct use of theory and to the validity of the PDE. Section 4.4 studies more specifically and more empirically one of the difficulties relating to the pricing of contingent claims on real estate, and tries to estimate the distribution of error. To this end a simplified reverse mortgage is presented and the impact of the appraisal accuracy (for the house associated to this contract) on the put option embedded in the mortgage is estimated. Data provided by IPD-France are used for this purpose. The conclusion presents briefly other possibilities for a financial modelisation of real estate, more connected to the reality of real estate assets, and discusses the overall possibility of pricing contingent claims on properties.

4.2. Applying financial theory to real estate : Optional models for mortgage

4.2.1. Literature review

MBS markets are an important field of investment in the United States. Mortgages entitled to individuals are grouped together, and representative bonds backed on this pool are sold to investors. This configuration gave rise to a lot of research on mortgage valuation because of the associated risks in the mortgage, namely the prepayment and default risks. Indeed, when mortgage rates drop, a wave of repayments arises, changing the financial characteristics of the MBS (mortgage proceeds are directly passed to bondholders, shortening the duration of their portfolios). This sensitivity to the rates led to a development of models using stochastic techniques[19]. We can mention among the first major contributions Dunn and McConnell (1981), who chose to work with one state variable (a CIR process), whereas Schwartz and Torous (1989) chose two processes for the short and long term rates. In both cases the models gave a partial differential equation (PDE) which must be solved numerically, its complexity preventing a closed form formula. Prepayment was described by using exogenous functions (a Poisson process in the first article, a proportional-hazard model in the second), which were calibrated on historical data and then incorporated into the PDE before its numerical implementation. Prepayment was not thought of as structural, it is not endogenous to the model. There was neither a house process in these first models which can be probably viewed as a gap as suggested by Downing, Stanton and Wallace (2003).

[19] Stochastic calculus and arbitrage theory are, since the 70's, a very powerful tool to deal with uncertainty in finance. Pioneers were Black and Scholes (1973) and Merton (1971, 1973) with their works on options pricing and optimal rules for consumption and investment. Diffusion processes allow handling easily expected returns and risk premiums in a continuous way, giving a rigorous framework for uncertainty.

There exist two levels of analysis for mortgages, so called loan level and pool level; the articles mentioned previously were focused on the second one. The behavior of the MBS is probably harder to study because of the portfolio structure and the interactions between loans gathered in the pool. For the loan level, Titman and Torous (1989) is one of the first attempts. But the numerous papers written by Kau, Keenan et al. are the reference for the valuation of mortgages considered as such and not as a part of a pool. We can mention their 1992 article for fixed-rate, and their 1993 for floating rate, while their 1995 survey is also a very interesting. The aim is to reach a sufficiently precise and flexible procedure for the numerical valuation of mortgages and of its components (prepayment and default options). To this end two processes are used, a CIR for the interest rate r and a geometric Brownian motion for the real estate value H. The mortgage is seen as a time-varying contingent claim relying on r and H, V(t, r, H) - the latter being the value of the mortgage -, and usual arbitrage theory is applied resulting to a PDE satisfied by V. Numerical results are obtained with a backward procedure, imposed by the forward looking behavior of borrowers in their exercise of the options. Calculations are made discretizing time and applying a finite difference algorithm on each interval with boundary conditions. Azevedo-Pereira et al. (2000) provide a good presentation for the numerical details of this method. A crucial point in this model is the endogenous aspect of the option exercise, contrary to the first articles mentioned above. Roughly speaking a prepayment occurs when refinancing mortgage rates are sufficiently low, and a default takes place when the borrower is in a situation of negative equity. It's no longer an exogenous treatment; the choice is analyzed into the structure of the model.

Thereafter, some improvements have been developed within this framework. Hilliard, Kau and Slawson (1998) produced a less cumbersome numerical procedure working with a bivariate binomial lattice. The default option had been split in two by Ambrose and Buttimer (2000) in order to analyze the behavior of deficient loans (default is not a one step process; it comprises two options, a right to stop making payments temporarily and a right to give up the property). Kelly and Slawson (2001)

introduced the effect of prepayment penalties in the borrower decisions when they exercise. Stanton (1995) improved the realism in the description of the borrowers' choices, assuming that they do not constantly reexamine their options but only at random moments, producing in the same time a convincing explanation for the burnout phenomenon. Option models had also been used as tools for pricing CMOs. For example McConnell and Singh (1994), incorporating Stanton's ideas, have been able to find a price. Ambrose, Buttimer and Thibodeau (2001) explained a part of the spread between Jumbo loans and conforming loan, using these theoretical models and the higher volatility observed on high-priced houses. Azevedo-Pereira, Newton and Paxson (2002 and 2003) fit this theory to specific mortgage products used in the UK. All these developments and adaptations have been integrated into the options framework without great difficulties, bringing to the fore the power and the applicability of this approach (see Kau and Slawson (2002) for one of the latest version of these models).

The purpose of what follows is to present the standard no-arbitrage argument producing the well-known valuation PDE, for the specific situation studied here, with one rate process and one real estate process[20]. The financial assumptions needed to establish this PDE will be pointed out as explicitly as possible, and their validity will be re-examined in the next section, which relies on this theoretical and quite classical first part.

4.2.2. Market and notations

In these articles the economy is described by two state variables: a spot rate r and a value of real estate H. Usually r follows a CIR process, H is a geometric Brownian motion, however we won't specify here explicitly the dynamics, calculations will be done using general diffusion processes.

[20] There exist numerous books dealing with arbitrage theory, we can refer for instance the Björk's, published by Oxford University Press: "Arbitrage theory in continuous time".

Under the objective probability we have:

$$dr = \mu_r (t, r, H)\, dt + \sigma_r (t, r, H)\, dz_r = \mu_r\, dt + \sigma_r\, dz_r \qquad (17)$$

$$dH = \mu_H (t, r, H)\, dt + \sigma_H (t, r, H)\, dz_H = \mu_H\, dt + \sigma_H\, dz_H \qquad (18)$$

$$dz_r\, dz_H = \rho(t, r, H)\, dt = \rho\, dt \qquad (19)$$

The drift for H doesn't include the service flow resulting from the possession of the house over time; it only describes the capital appreciation of the real estate.

A prepayment option, a default option or a mortgage taken in its whole (i.e. a loan with the two options) are contingent claims relying on time, r and H, we denote it by V(t, r, H). We will assume, as is required by arbitrage theory, that V is perfectly tradable in a frictionless market. That hypothesis is very strong but essential to get the PDE valuation.

V(t,r,H) depends on the two state variables and is a diffusion process too, and Itô's lemma gives its dynamics :

$$dV = [V_t + \mu_r V_r + \mu_H V_H + \tfrac{1}{2} V_{rr} \sigma_r^2 + \tfrac{1}{2} V_{HH}\, \sigma_H^2 + V_{rH}\, \rho \sigma_r \sigma_H]\, dt + \sigma_r V_r dz_r + \sigma_H V_H dz_H$$

For simplicity we denote the drift DV (D for Dynkin)

Therefore $dV = DV\, dt + \sigma_r\, V_r\, dz_r + \sigma_H\, V_H\, dz_H$

Dividing by V $dV / V = a\, dt + s_r\, dz_r + s_H\, dz_H$

Where $a = DV / V$ $s_r = (\sigma_r\, V_r) / V$ and $s_H = (\sigma_H\, V_H) / V$

In order to establish the PDE we have to select two particular assets $V_1(r,H,t)$ and $V_2(r,H,t)$, perfectly tradable in a frictionless market as well. They will serve as

primary assets, allowing the construction of a riskless portfolio. With the same notation we have:

$$dV_i / V_i = a^i\, dt + s^i_r\, dz_r + s^i_H\, dz_H \qquad i = 1, 2 \qquad (20)$$

4.2.3. Riskless portfolio and PDE

An investor chooses to build a portfolio, buying the securities V, V_1 and V_2 in quantities N, N_1 and N_2. This idea is not purely formal; the portfolio must be practically possible.

We denote W his total wealth, $W = N V + N_1 V_1 + N_2 V_2$ (21)

Differentiating $dW = N\, dV + N_1\, dV_1 + N_2\, dV_2$

$$\begin{aligned} &= N V (a\, dt + s_r\, dz_r + s_H\, dz_H) \\ &\quad + N_1 V_1 (a^1\, dt + s^1_r\, dz_r + s^1_H\, dz_H) \\ &\quad + N_2 V_2 (a^2\, dt + s^2_r\, dz_r + s^2_H\, dz_H) \\ &= (NV a + N_1 V_1 a^1 + N_2 V_2 a^2)\, dt \\ &\quad + (NV s_r + N_1 V_1 s^1_r + N_2 V_2 s^2_r)\, dz_r \\ &\quad + (NV s_H + N_1 V_1 s^1_H + N_2 V_2 s^2_H)\, dz_H \end{aligned}$$

If we can find N, N_1, N_2 verifying:

$$\begin{cases} NV s_r + N_1 V_1 s^1_r + N_2 V_2 s^2_r = 0 \\ NV s_H + N_1 V_1 s^1_H + N_2 V_2 s^2_H = 0 \end{cases} \qquad (22)$$

the corresponding portfolio will be riskless and its drift, using the no-arbitrage assumption, will be : r W.

For example if $s^1_r\ s^2_H - s^1_H\ s^2_r$ is different from zero, for each value of N we can find N_1 and N_2 solving this system. As this portfolio is riskless we can then assert that (N, N_1, N_2) is also a solution for the following system :

$$\begin{cases} NV\,a + N_1V_1\,a^1 + N_2\,V_2\,a^2 = r\,W \\ NVs_r + N_1V_1\,s^1_r + N_2\,V_2\,s^2_r = 0 \\ NVs_H + N_1V_1\,s^1_H + N_2\,V_2\,s^2_H = 0 \end{cases} \qquad (23)$$

Writing $W = N\,V + N_1\,V_1 + N_2\,V_2$, we get:

$$\begin{cases} NV(a-r) + N_1V_1(a^1-r) + N_2V_2(a^2-r) = 0 \\ NV\,s_r + N_1V_1\,s^1_r + N_2\,V_2\,s^2_r = 0 \\ NV\,s_H + N_1V_1\,s^1_H + N_2\,V_2\,s^2_H = 0 \end{cases}$$

The initial wealth W not being zero (N, N_1, N_2) cannot be the null vector. As the above system admits a non trivial solution its determinant must be zero. Transposing it, it gives:

$$\begin{vmatrix} a-r & s_r & s_H \\ a^1-r & s^1_r & s^1_H \\ a^2-r & s^2_r & s^2_H \end{vmatrix} = 0$$

The rules of linear algebra imply, then, a linear dependence relation between the columns. For instance the first one can be written as a combination of the second and the third.

In others words, there exist $\lambda_1 = \lambda_1(t, r, H)$, $\lambda_2 = \lambda_2(t, r, H)$ such that :

$$\begin{cases} a - r = \lambda_1 s_r + \lambda_2 s_H \\ a^1 - r = \lambda_1 s^1_r + \lambda_2 s^1_H \\ a^2 - r = \lambda_1 s^2_r + \lambda_2 s^2_H \end{cases} \qquad (24)$$

The two last equations are not depending on the features of V, therefore by requiring that $s^1_r\ s^2_H - s^1_H\ s^2_r$ is different from zero we can express λ_1, λ_2 independently of V. The fact that these quantities are the same for all assets V(t,r,H) is a major result and will lead directly to the notion of market price of the two sources of risk.

Regarding the first equation, it gives the PDE:

$$a - r = \lambda_1 s_r + \lambda_2 s_H$$

$$DV - rV = \lambda_1 \sigma_r V_r + \lambda_2 \sigma_H V_H$$

Replacing DV with its expression:

$$V_t + \mu_r V_r + \mu_H V_H + \tfrac{1}{2} V_{rr} \sigma_r^2 + \tfrac{1}{2} V_{HH} \sigma_H^2 + V_{rH} \rho \sigma_r \sigma_H - rV = \lambda_1 \sigma_r V_r + \lambda_2 \sigma_H V_H$$

Rearranging it, V(t, r, H) is thus a solution of the following PDE :

$$V_t + (\mu_r - \lambda_1 \sigma_r) V_r + (\mu_H - \lambda_2 \sigma_H) V_H + \tfrac{1}{2} V_{rr} \sigma_r^2 + \tfrac{1}{2} V_{HH} \sigma_H^2 + V_{rH} \rho \sigma_r \sigma_H = rV \qquad (25)$$

4.2.4. Interpretation of λ_1

There are choices more natural than others for the two benchmark assets; namely V_1 which only depends on r, and V_2 which only depends on H (for example V_1 could be the money market account, H could be V_2 , but subject to the inclusion of the service flow produced by the real estate during its holding period). With such a choice, the mathematical formula becomes simpler and the interpretation easier.

As we have $V^1_H = s^1_H = 0$ and $V^2_r = s^2_r = 0$ system (22) becomes:

$$\begin{cases} NVs_r + N_1 V_1 s^1_r = 0 \\ NVs_H + N_2 V_2 s^2_H = 0 \end{cases} \qquad (22')$$

And the last two equations in (24) are now:

$$\begin{cases} a^1 - r = \lambda_1 s^1_r \\ a^2 - r = \lambda_2 s^2_H \end{cases}$$

Which means that :

$$\begin{cases} \lambda_1 = (a^1 - r) / s^1_r \\ \lambda_2 = (a^2 - r) / s^2_H \end{cases} \qquad (26)$$

For any asset only relying on t and r, the first equation in (24) gives :

$$a - r = \lambda_1 s_r$$

hence :

$$\lambda_1 = (a^1 - r) / s^1_r = (a - r) / s_r \qquad (27)$$

These two quotients represent the market price, per unit of volatility, for the risks associated with the assets V(t, r) and V_1. The dynamics of V(t, r) and V_1 (t, r) are $dV/V = a\,dt + s_r\,dz_r$ and $dV_1/V_1 = a^1\,dt + s^1{}_r\,dz_r$; (a – r) and (a^1 – r) are the risk premiums and s_r , $s^1{}_r$ the corresponding volatilities. This equation establishes that the risk price is always the same, whatever the asset (if the latter relies only on t and r); it is a relation of internal coherence following from the no-arbitrage assumption. λ_1 is an exogenous process, it is determined by the market, and once its price is known the drift of any asset V (t, r) is $a = r + \lambda_1 s_r$ (the riskless rate increased by the risk contribution coming from the volatility).

4.2.5. Risk neutrality

Dynamic of r is : $dr = \mu_r\,dt + \sigma_r\,dz_r$

Introducing λ_1 : $dr = (\mu_r - \lambda_1 \sigma_r)\,dt + \sigma_r\,(dz_r + \lambda_1\,dt)$

The same for H : $dH = \mu_H\,dt + \sigma_H\,dz_H$

$dH = (\mu_H - \lambda_2 \sigma_H)\,dt + \sigma_H\,(dz_H + \lambda_2\,dt)$

Assuming that λ_1 and λ_2 are known, we can use multidimensional Girsanov theorem. There exists a measure Q under which dynamics of r and H are:

$dr = (\mu_r - \lambda_1 \sigma_r)\,dt + \sigma_r\,d\check{z}_r$

$dH = (\mu_H - \lambda_2 \sigma_H)\,dt + \sigma_H\,d\check{z}_H$ (with $d\check{z}_r\,d\check{z}_H = \rho dt$)

And, since V is a solution of the PDE:

$$V_t + (\mu_r - \lambda_1 \sigma_r) V_r + (\mu_H - \lambda_2 \sigma_H) V_H + \tfrac{1}{2} V_{rr} \sigma_r^2 + \tfrac{1}{2} V_{HH} \sigma_H^2 + V_{rH} \rho \sigma_r \sigma_H = r$$

in which the terms in front of V_r and V_H are, this time, the same as the drift for the dynamics of r and H under Q. We can then apply Feynman-Kac theorem obtaining the fundamental formula:

$$V(t,r,H) = \mathbf{E_Q}\left[e^{-\int r(u)du} V(T,r(T),H(T)) \right] \quad \text{for } t \leq T, \text{ integral between t and T}$$

Discounted price of perfectly tradable assets are martingales under Q.

4.2.6. H is a tradable asset, interpretation of λ_2

The real estate (physical house and service flow) can be bought and sold. If we consider it can be traded in a perfect and frictionless market, its value is then determined by a PDE solution. Let us note H^* the house process including these "dividends", we have:

$$dH^* = dH + \delta(H)\, H\, dt \qquad (\delta(H) \text{ is similar to a convenience yield})$$
$$dH^* = (\mu_H + \delta(H)\, H)\, dt + \sigma_H\, dz_H \qquad (\text{under the objective measure})$$

And under Q : $dH^* = dH + \delta(H)\, H\, dt$

$$dH^* = (\mu_H - \lambda_2 \sigma_H + \delta(H)\, H)\, dt + \sigma_H\, d\check{z}_H$$

H^* being a martingale under Q, its return is also r :

$$\mu_H - \lambda_2 \sigma_H + \delta(H)\, H = r\, H^* = r\, H$$

($H^* = H$ doesn't create an arbitrage opportunity because H is not a tradable asset, it is only an abstract process giving the price of the real estate at t)

Thus we have: $\mu_H - \lambda_2 \sigma_H = (r - \delta(H))\, H$

Or else: $$\mu_H + \delta(H)\, H = r\, H + \lambda_2\, \sigma_H \qquad (28)$$

The left hand side is the instantaneous return under the objective measure, associated with the owning of H^*, the equality splits this quantity in a riskless part, rH, and $\lambda_2\, \sigma_H$. λ_2 is then the risk premium per unit of volatility for H and more generally for all the assets only relying on H. It's the market price for the house risk.

4.2.7. Summing-up

For a contingent claim V(t, r ,H), depending on state variables r and H we can calculate the no-arbitrage price solving the following PDE :

$$V_t + (\mu_r - \lambda_1\sigma_r)V_r + (rH - \delta(H)H)V_H + \tfrac{1}{2}V_{rr}\, \sigma_r^2 + \tfrac{1}{2}V_{HH}\, \sigma_H^2 + V_{rH}\, \rho\sigma_r\sigma_H = rV \qquad (29)$$

Where :

- $\mu_r - \lambda_1\, \sigma_r$ equals the drift of r under the risk neutral measure Q. Usually the dynamic of r under Q is described using a CIR process ($dr = \gamma (\theta - r)\, dt + \sigma_r\, r^{1/2}\, d\check{z}_r$), so $\mu_r - \lambda_1\, \sigma_r$ must be replaced with $\gamma (\theta - r)$ in the equation
- $\delta(H)$ is the service flow coming from the house, usually assumed constant
- σ_r , σ_H , ρ are the volatilities and the correlation for r and H under both measure

Very often V is a prepayment or a default option. Numerical solutions are calculated using backward procedure and finite difference methods specifying some boundary conditions (see Azevedo-Pereira (2000) for more details).

4.3. From theory to practice

The previous demonstration ends up to equation (29). It is of course a mathematical object but the strong assumptions made during the process are primarily financial. After this technical step we are going to question these hypotheses in the light of the real estate markets. If they are not empirically verified, the mathematical demonstration, as sophisticated as it may be, will only produce an abstract and biased formula.

4.3.1. The no-arbitrage assumption

The main result established above is the possibility of valuing an option embedded in a loan, secured by a real estate, using numerical methods for PDE. This fair price is a consequence of a market mechanism, formalised by the no-arbitrage assumption. Indeed, at any moment, if V(t, r, H) moves away from the theoretical value, it would imply the existence of an arbitrage opportunity. Arbitrageurs would then take advantage of the situation and the attempts to realize the potential benefits would bring back V to its theoretical level.

In the real world, bankers are not spending all their time solving PDEs and the job of traders is not the application of finite difference methods. The no-arbitrage assumption must be understood as an indirect result of the market activity. If someone detects a local inefficiency, he will act to take advantage of it, but in the majority of times this will be done without the use of mathematics. For example if the exchange rate is 0.6$ for 1€ in a market and at the same time 1€ is valued 1.4$ on another market, one does not need to solve a PDE to earn a lot of money. On the total market scale the actions of all participants lead to the no-arbitrage assumption; this hypothesis is supported by an actual and concrete activity. It allows us to assert that a riskless portfolio will have a return equal to the riskless rate, and it produces the

valuation equation[21]. The PDE is then an indirect consequence of the local arbitrages realised by the market participants and for it to be valid some precise and important conditions must be fulfilled.

4.3.2. Real estate markets

Certainly the stock market operates in a way that is much closer to the assumptions of arbitrage theory than the real estate market. Usually, when studying a problem, the first approach takes an ideal situation in order to establish closed and quite simple formulas, while further approaches refine – and possibly change - the initial hypotheses in order to achieve a more realistic description. For instance the Black-Scholes formula had been initially established assuming constant volatility, thereafter this assumption had been relaxed (for example Heston (1993)). Nevertheless the foundations of the arguments were maintained and the situation remained to a great extent unchanged.

For real estate this stage of improving on the initial simple assumptions is not at all straightforward. Problems can arise and a blind application of conventional techniques can produce misleading results. For example we saw previously that the PDE relies on taking action of the opportunities of arbitrage, and that it is not just an abstract hypothesis. Consider the case of a prepayment option, the previous argument would mean that there actually exist people that are realising the arbitrage opportunities between V_1, V_2 and the prepayment option. These traders should buy or sell houses, lend or borrow, in order to make money with the small inefficiencies in the mortgage contracts. Of course it is an unrealistic situation and this kind of actions does not take place. Options embedded in mortgage contracts have prices which depend more on competition between lenders rather than on hypothetical arbitrages. The arbitrage activity required in theory is thus absent for a prepayment option.

[21] This assumption is used when we get (23) from (22)

The arbitrage based modelling also assumes implicitly that assets can be isolated and priced separately. However for residential loans it is not always as easy as it seems because of the cross-selling problem. For investors this matter does not exist when the mortgages are securitized in an SPV, since the only proceeds they will receive will be the passed monthly payments, and the pool will be in a "situation of isolation". But when a bank chooses to keep the loans in its balance sheet this assumption of isolation becomes questionable. In France for example, residential loans are essentially loss makers; expected earnings for lenders are not only coming from the contract. They also come from other financial products that the bank can sell to its customers, as for example a current account or a consumer credit, taking advantage of the privileged relationship established between lender and borrower during the loan process. Thus if a bank tries to value a mortgage using the PDE, cash flows should be modified in order to take into account the other indirect earnings[22]. The problem will then be to find a precise valuation method for these "blurred" expected proceeds.

The use of SPV or other isolation techniques such as covered bonds makes easier the financial study of the mortgage; interferences with other asset classes are low or even non-existing in the perfectly isolated cases. But in both cases, the portfolio structure must also be examined for the mortgages class. The interactions between the mortgages inside the SPV or inside the balance sheet and their possible correlations are also a matter of importance.

Summing up, prepayment option can sometimes be a commercial problem rather than a financial one, assets are never isolated and real arbitrages are non-existent. This questions the validity of the arbitrage pricing. In the following three paragraphs we will see precisely where the difficulties are in the actions (required by theory) of taking advantage of the arbitrage opportunities, for the benchmark assets V_1 and V_2 as for the contingent claim V.

[22] Securitization is strongly developed in Anglo-Saxon countries, partly explaining the greater success and the greater use of the models presented above since the isolation hypothesis is quite realistic.

4.3.3. Asset associated with the rate risk

In the first section, V_1 indicated a time-varying asset relying only on r; it was the money market account. Interest rate products are sufficiently standardised, negotiated and liquid. Thus we can reasonably consider V_1 as a perfectly tradable asset in a frictionless market; it can be bought or sold without any limitations. Building an arbitrage strategy based on lending or borrowing money is not a problem. Of course one needs to make a choice on the specific model for the risk neutral process for r. Which should be used between Vasicek, CIR or Hull-White? However this question is not specific to the problem examined here. Since there already exists an established theory for the pricing of interest rate products; we are going here to utilise its results, not questioning further its validity.

4.3.4. Housing benchmark

In the arguing made in the previous section, V_2 was not specified explicitly and mathematically we could choose a quite complex asset depending on t, r and H : $V_2 = V_2(t, r, H)$. The only thing we needed to ensure to solve system (22) was that $s^1_r \; s^2_H - s^1_H \; s^2_r$ was different from zero, and that condition is not a very demanding one. Among all possible choices, there are some that are better than others in the sense of them providing better intuition. Since V_1 is a benchmark for r, V_2 should also be a benchmark for H. Interpretation would be greatly facilitated by $V_2 = V_2(t,H)$

The most natural choice for V_2 is H^* itself (the value of the house securing the loan), and as H^* is a tradable asset, we could apply the theoretical results easily. Indeed, in arbitrage theory when a state variable is not traded, problems can arise. For example models with stochastic volatility introduce difficulties because of the impossibility of trading directly the volatility for hedging a portfolio. Thus $V_2 = H^*$ should be preferable for modelling. However, in this case also, one can have objections stemming from the feasibility of financial operations. The latter should be

possible practically and not only in a perfect world (as described by arbitrage theory).

A first difference with the theoretic situation comes from the uniqueness of the house. If someone tries to build an arbitrage, on a prepayment option, he cannot own the specific house associated to the mortgage because it is already the borrower's ownership. He can only try to build his riskless portfolio with a similar real estate, adding, unfortunately, a risk of imperfect replication. Houses cannot be substituted one with another as two stocks can. A second problem may occur if the strategy implies the sell of a house. If our hypothetical trader does not have it already in his portfolio, it can be difficult to take a short position on a house. And last but not least, H^* is unknown until the sale completion. Property prices are not quoted as stocks, the price of a real estate is only revealed in a transaction. It is very far from the "prefect" financial situation assumed by arbitrage theory, making the construction of a riskless portfolio practically impossible. Several articles written by Childs et al. (2001-2002a-2002b-2004) deal with this problem, analysing H^*as a noisy asset, as for example in real options theory. Considering H^* as financial asset would bring out the appraisal problem. As the right price cannot be known without a transaction, the only means to carry out a financial analysis is by forming an estimation; its quality becoming then a central point. Usually estimation is considered correct when the spread between the real price and the estimated price is less than 10%. We will see later what can the consequences be of the appraisal problem on a pricing example. Therefore a rather natural choice for V_2 is strongly rejected for purely practical reasons.

Another possibility for V_2 could be a REIT stock. This provides an important improvement compared to the previous choice, namely the liquidity. These assets can be bought or sold very easily and prices are known at every time. Nevertheless there are always important difficulties. For example which REIT is the more appropriate? In addition, REITs are usually correlated with other stocks in the market, resulting to the introduction of additional (market) risk. In terms of the model developed here, it means the introduction of a third Brownian motion, complicating the building of

riskless portfolio. Theoretically two benchmark assets with three random sources do not allow the building of a risk-neutral portfolio whatever the combination of V_1 and V_2. In other words, an interesting and practical choice for V_2 leads to a mathematical problem.

Finally the last possibility we are going to consider for V_2 is an index (for instance a price/m² on a determined area for residential property). The liquidity problem appears here also because an index is not a directly tradable asset. If the requirements for hedging or for building a risk-neutral portfolio include a purchase of only 3m² (three times the index) of residential property in Paris, they could be hard to realise. However a natural and interesting way to achieve this could be the use of an index derivative. It would suppose a quite mature and liquid market for such products, but at the present time it is not a reality. A possible future development of the market for index derivatives could lead to a greater use of arbitrage theory in real estate, which is certainly a promising possibility.

In conclusion, the choice of V_2 is not obvious. The financial features of a house are not in good agreement with the requirements of arbitrage theory, REIT stocks bring another random source and the indexes are not enough developed. It seems rather inappropriate not to take this reality into account when using theoretical models.

4.3.5. Contingent claim V

PDEs are essentially applied to prepayment and default options. Once more it means that these assets are perfectly tradable in a frictionless market and unfortunately, in this case too, reality is far from this. We saw before that traders cannot take advantage of possible arbitrage opportunities on these options, resulting in their prices being defined more by bank policies than by market actions. Nevertheless, if someone wanted to take advantage of a hypothetic arbitrage opportunity, s/he will be

faced with another practical difficulty coming from the insufficient financial features of these options. To be specific, these options are embedded in the mortgages. It is therefore impossible to own or to sell an option separately from the loan. The consequences of this concern system (22). A riskless portfolio requires the finding of integers N, N_1, N_2 making the random contributions of the two Brownian motions zero. A mathematical solution could perfectly be $N_1 = N_2 = 1$ and $N = 4$; one unit of the two benchmarks assets and four units of a prepayment option. Who could sell separately these four options? If one tries to buy them indirectly, by signing into four mortgages, s/he would result in adding others risks linked to the loans; consequently it would no longer be a riskless portfolio. A mathematical arbitrage opportunity is not always a financial one since the options are not isolated securities.

4.3.6. Validity of the PDE

Real estate is not as smooth a financial product as stocks. Its peculiarity is illustrated by the inability of traders to take advantage of possible arbitrage opportunities. However the problem is not actually produced by this inability, it is really the reasons behind the inability that are responsible. The reason why arbitrage activity in real estate is scarce, is mostly because it is practically very problematic to construct a riskless portfolio. Among the reasons for such practical problems one can mention the uniqueness of house, the impossibility of being short on a real estate, the price uncertainty until a real transaction is made, the introduction of market risk by the use of REIT, the underdevelopment of index derivative markets and the implicit nature of options that cannot be traded on their own.

The modelisation developed in section 4.2 ends up with a valuation PDE which can be implemented using numerical procedures. It is a very powerful result but fundamentally it assumes some very strong conditions on the securities under study. In a sufficiently well developed, mature, liquid and frictionless market (in other words in a “perfect” financial market) the no-arbitrage assumption is acceptable.

This assumption is the core of the model and it allows us to assert that a riskless portfolio must give a return of r (23), producing finally the equation (25). In this section we have just seen the situation in the real estate market is far from this. The imperfections involved are not secondary difficulties which could be resolved by slightly changing the assumptions, as it was done with stochastic volatility models. The imperfections here question the model on the whole, and ultimately, the accuracy and validity of the final PDE. The modelling is not sufficiently linked to real estate particularities, and a direct application of arbitrage theory, originally built for "perfect" financial products such as stocks, can produce misleading valuations.

4.4. Example of a misleading pricing

Among the difficulties previously mentioned, we are going to elaborate on one of them; namely the difference between the right price and the estimated price. The aim is to get an idea of the size of the potential pricing error for contingent claims written on real estates. In particular we will examine a simplified version, namely the put option embedded in a reverse mortgage.

4.4.1. Asset description

Recently a report (Jachiet et al. 2004) ordered by the French government has studied the possibility to create, in France, products that make easier the availability of residential equity that has accumulated in elderly persons' houses. These assets are already known in UK as lifetime mortgages and in the US as reverse mortgages; they are examples of the home equity loans family (loans granted for consumption and warranted by real estate). In France this type of product, named "crédit hypothécaire", is scarce, but there exist governmental projects aiming at creating and developing this possibility (the "prêt viager hypothécaire", which translates to "reverse mortgage").

The principle of the contract is as follows: At the origination the lender provides the borrower with certain amount. During the following years interests are added to the initial amount lent, without any repayments required from the borrower. At the event of death of the borrower, the house warranting the loan is sold and the bank is repaid. Two situations can then occur: the sale price is higher or lower than the accumulated debt. In the former situation the difference goes to the heirs of the borrower, and in the latter, a clause included in the initial contract limits the bank's rights to repayment up to the house's sale price. This creates a situation where the bank has effectively shorted a put option on the value of the house. The major risks concern the life expectancy and the final house price, it is a contingent claim depending on r and H[23] : V(t, r, H). The purpose here will not be to examine rigorously all the features of this loan and to calculate the right price of such a contract[24]. We are going to work on a simplified version, in order to understand the consequences for the pricing of the put option and the mortgage, of the fact that the value of the real estate is not explicitly known. Indeed, at the beginning, H^* is not known directly as long as the house is not sold, giving rise to an error in the estimation of the house value, which leads directly to a mispricing for the assets relying on H^*. Accurate estimations are always have always been and continue to be important for real estate finance.

4.4.2. Simplified contract and environment

Let us assume that at t = 0, the lender pays K to the householder. Interest is accruing continuously at a fixed annual rate r, and, for simplification, let us assume that the rates curve is flat and does not change with time. We assume further that the contract is executed at a fixed date T corresponding to the life expectancy of the borrower (for example ten years). Notice here that this view is not very realistic, but it avoids the

[23] Here, H represents the real house price process
[24] It would require the use of an American option and the modelisation of the death of the borrower as a stopping time. The tools coming from life insurance would be relevant to this end.

complications of working with what is called a stopping time, associated with the random time of death of the borrower. Moreover a fixed maturity allows the use of a closed formula for the price of a European put option.

At t = T, the lent capital has produced a debt of $Ke^{rT} = K'$, however this amount is capped at H(T), so the final payoff for the bank is :

$$\text{Min} (K' ; H(T)) = K' - \max (0, K' - H(T)) \qquad (30)$$

In other words, the bank's portfolio contains a loan and a short position on a put. This option is European, written on H, with maturity T and strike of K'.

We assume that the dynamics of H are given by a geometric Brownian motion

$$dH = \mu H \, dt + \sigma H \, dz$$

Hence we can use the Black-Scholes formula to value this position. At t = 0, as $e^{-rT} K' = K$ the worth of the portfolio is:

$$K - (e^{-rT} K' N(-d_2) - H(0) N(-d1)) = K - \text{Put}(0) \qquad (31)$$

Where : $d_1 = \{ \ln(H(0) / K') + (r + \sigma^2/2) T \} / \sigma T^{1/2}$

and $d_2 = d_1 - \sigma T^{½}$ (N being the standard normal distribution function)

Uncertainty in this formula enters with the term H(0), since the real value cannot be observed exactly without a sales transaction. The next calculation is made assuming that H(0) is known, hence we can get the true prices of the contingent claims.

Choice of parameters:

- H(0) = 100 000 (real price of the house)
- T = 10 years [25]
- K = 70 000
- r = 5%
- K' = Ke^{rT} = 115410,49
- μ = 2% [26]
- σ = 15% [27]

At T, the final debt will be K' = 115410,49 and if $\sigma = 0$, the final value of the house will be H(0) $e^{\mu T}$ = 122140,28. In the absence of uncertainty the debt would not exceed the value of the sale, and the heirs of the borrower would receive the difference. If $\sigma \neq 0$, the Black-Scholes formula is applied, and in this situation of full information the true put value is Put(0) = 5101,92. The mortgage price is 70 000 – Put(0) = 64898,08.

It must be noticed that the calculation of Put(0) implicitly assumes that the arbitrage assumptions made in the Black-Scholes formula are valid in the case of a real estate derivative. In particular that the riskless portfolio exists and traders are able to take advantage of arbitrage opportunities (we saw previously what the problems are in making such assertions). Here, the aim being to examine the consequences of one of these problems, namely the price uncertainty on H(0), the options will be priced assuming that real estates are perfectly tradable in a frictionless market. We will work as if the difficulty with the appraisal was the only remaining imperfection. We could hope that with such a nearly "perfect" context things would go quite well, but unfortunately this uncertainty on H(0) makes, on its own, the pricing very problematical.

[25] Coherent with what is observed in the countries where the reverse mortgages exists. People subscribing are around 70 years old, a life expectation of 10 years is reasonable particularly for women

[26] μ isn't useful in the calculations of P(0). It's well known that options prices are independent of the drift; the only thing that matters in the dynamic of H for P(0) is the volatility.

[27] Coherent with the volatility estimated by Ambrose, Buttimer, Thibodeau (2001).

4.4.3. Incomplete information

For an appraiser, the estimation of H(0) can be considered correct if the error stays within a 10% interval of the real price. This accuracy is very reasonable when valuing such a complicated asset as a house, but unfortunately even such a good appraisal is going to produce a bad pricing for the put option. For example, if the house is valued at the origination with H (0) = 95 000, the price calculated for P(0) will be Put (0) = 5975,75. The option price will be overestimated by 17% and the mortgage value will be underestimated by 1,3% (estimated price = 64024,25). If the estimation gives H(0) = 110 000, we will have P(0) = 3722,78 (-27%) and a mortgage value of 66277,22 (+2,1%).

Let us now do a more detailed analysis in order to reach a better understanding of the phenomenon of such large percentage errors. The method developed thereafter does not aim at producing perfect estimations, it is rather rough and can admit a lot of improvements . The purpose is only to measure the size of the error when valuing the put. Let us assume that the percentage error on the estimated price is distributed according to a normal law $\mathcal{N}(m,s)$. The actual mean m is not necessarily 0 since there can exist a positive or negative bias. Drawing a sample $\{e_i\}$ distributed as described, we can then build a sample of estimated prices $\{ h_i = H(0) (1+ e_i) \}$[28]. For each h_i, the estimated option and mortgage prices can be calculated using a basic pricer. The differences between these values and the real ones provide two series, which we name put_error and mortgage_error, whose density can be represented using a kernel method.[29] For example, with m = 0 and s = 0,1 we obtain the following empirical distributions :

[28] If $h_i < 0$, h_i is replaced with 0.
[29] The size of the sample is chosen at 1000 to achieve a sufficient level of smoothing.

Figure 8 : distribution of error when estimating the house price (%)

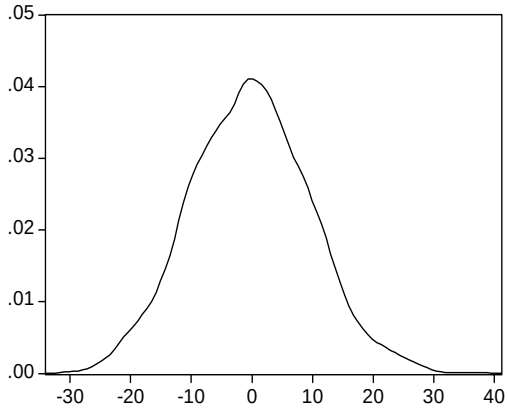


Figure 9 : distribution of put_error (%)

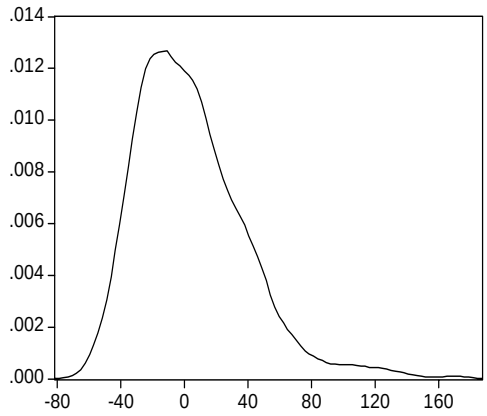
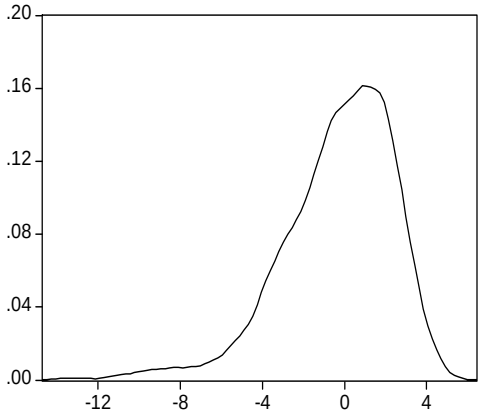


Figure 10: distribution of mortgage_error (%)



For this choice of parameters, the probability of having an error between -15% and +15% on the estimated house price is around 0,87. This type of situation is not at all unrealistic. However the probability of the event { |put_error| ≥ 40%} is near 20%, which means that on average every five estimations the put option is completely mispriced (the error is more than 40%). The noise on P(0) is an amplification of the noise on the pricing of H(0).

The mortgage comprises a put option whose real price is around 5000 and a loan valued 70000. Uncertainty exists only on the first part, and considering the relative sizes of these two components, the distribution of error becomes necessarily tighter. For instance, the event {|mortgage_error| ≥ 5%} has a probability of around 6%. However this – not bad - pricing is more a consequence of a size effect rather than a

well performing valuation methodology. The noise shrinks with the inclusion of the non-random component.

With a choice of m = 0 and s = 0.01, we have $P(|put\_error| \geq 5\%) \approx 11\%$. It could be considered as a reasonable pricing, but in order to achieve it, the accuracy of the house estimation should be very high, since the same parameters imply that $P(|house\ error| \leq 2\%) = 95\ \%$. In other words the appraisal in this case is quite always very effective; needless to say that such a situation hardly occurs in reality.

4.4.4. Error on P(0) when using real data

IPD-France publishes annually a property index, based on a set of real estate portfolios, highly representative of the total market. Appraisal is a central point for the reliability of this index since the majority of the assets are not sold each year. The quality of the assessments made by the experts becomes a matter of importance and it is studied in particular. Figure 11 shows the improvements in the appraisal quality, by showing the distribution of the spreads between valued and real prices when a sale is effectively realized soon afterwards the estimation.

Means and standard deviations can be roughly estimated with this histogram. For illustration the 27% that appears on the left axis and corresponds to the [0%;10%] interval for the year 1998, can be assigned the meaning of a 27% of the appraisals having been underestimated by 5% on that year. We use the value of the mid of each interval (in our example the [0%;10%] interval), hence the 5%. Results are in table 3 where we can see that there exists a slight trend to overestimate (negative mean). But the most important and notable characteristic is that the variance of the estimation is decreasing year by year, implying that estimations have greatly improved.

Figure 11 : difference between sale price and estimation 1998-2003, Source IPD-France[30] (left side: overestimated, right side: underestimated)

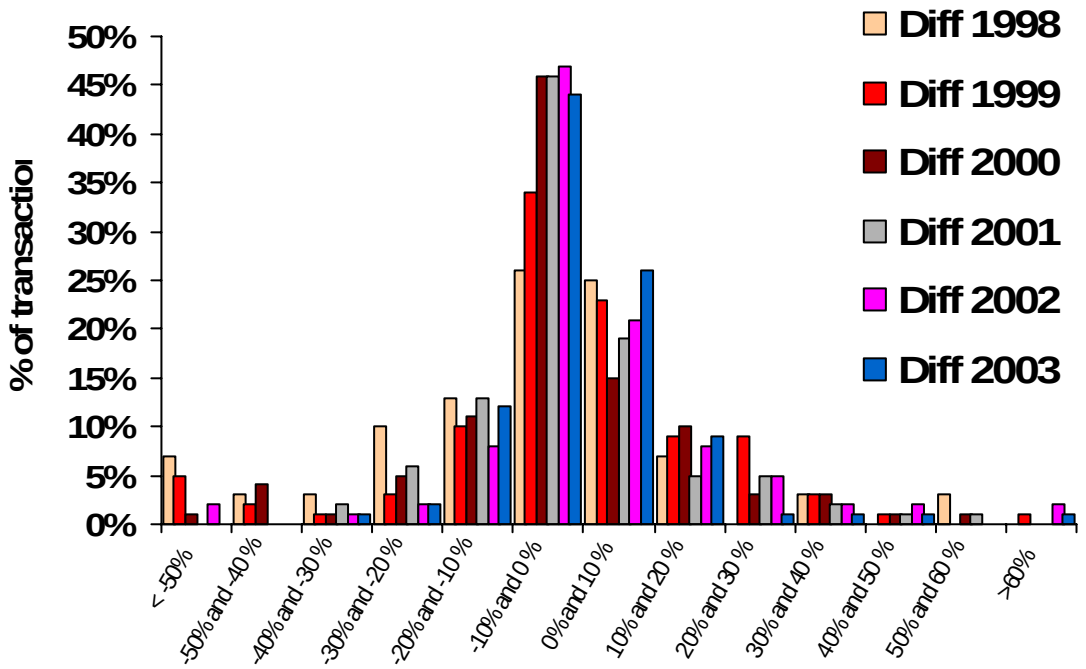


Table 3 : Estimation of mean and standard deviation for the distributions of Figure 11

	1998	1999	2000	2001	2002	2003
Mean	-0.070	-0.010	-0.028	-0.018	0.009	-0.005
Standard deviation	0.354	0.232	0.247	0.201	0.151	0.148

Assuming that the distributions are normal[31] with the parameters stated on the above table, we can use the same methodology as in the previous section. A sample is drawn from each year's distribution, and subsequently one can obtain the estimation for the house $\{h_i\}$, from which the put price is calculated with the Black-Scholes formula. Error on P(0) is measured for each house value and the shape of its distribution is obtained with a kernel density method. Figures 12, 13, 14, 15, 16, 17

[30] In the sample studied here, properties are not differentiated between retails, offices, industrial or residential. For a reverse mortgage data specifically focused on residential appraisal would have been more desirable; however we can reasonably expect that the differences between classes are not too important.

[31] This hypothesis is probably doubtful but here the matter is only to study the consequences of a decreasing in volatility on P(0).

present the results for the years 1998 to 2003.[32] The increase of the accuracy relating to the estimation of H(0) produces a tightening of the distribution around 0, in other words the put price is valued more accurately. This improvement is very significant when comparing for example figure 12 and figure 16; in the first case the error could sometimes be as large as 500%, and in the second it stays mostly under 100%. Having said that, even if in 1998 the situation was as bad as not being able to reasonably determine the put value, in 2002 and 2003 it still remains a difficult task. The noise on P(0) is always amplified compared to the noise on H(0); the usual leverage effect of the option brings about a very unpleasant consequence, acting as a noise amplifier. If we had studied directly a pool of mortgages, the global appraisal noise would have been lower, but unfortunately what we could have got on one side would have been lost on another. Indeed, a modelling of the correlations is required when analysing a portfolio, and the choice of a geometric Brownian motion for H becomes then much more problematic.

Summing up, a financial approach of real estates brings the quality of appraisal at a central place. However for some claims as options the noise effect prevents a traditional pricing. This difficulty seems to be rather important, since $P(|\text{put\_error}| \geq 5\%) \approx 11\%$ is associated as mentioned earlier with $P(|\text{house error}| \leq 2\%) = 95\ \%$. In other words, the put price is correctly determined only if there is almost no appraisal error. Good quality of appraisal is not a guarantee for precise pricing, particularly for leveraged assets as options. More generally, the usage of very sophisticated methods such as PDEs does not guarantee an equal standard of pricing accuracy, in this case the sophistication of methodology is rather disproportional to the quality of results. A rough estimation of the prepayment option, for example 3% of the remaining capital balance, is of course inexact, but it is not obvious that the PDE estimation is better. What's more, the former price is very simply obtained (because it is determined by experience), while the latter is not

[32] Observing the graphs we can notice that the probability of having an error on P(0) inferior to -100% is not null; it means that the option prices can be negative. This imperfection in the distribution comes from the kernel method; a better estimation would truncate this part.

at all, and that is obvious. In a costs / benefits analysis, the PDE estimation is not fully convincing.

Figure 12: distribution of put_error_1998 (%)

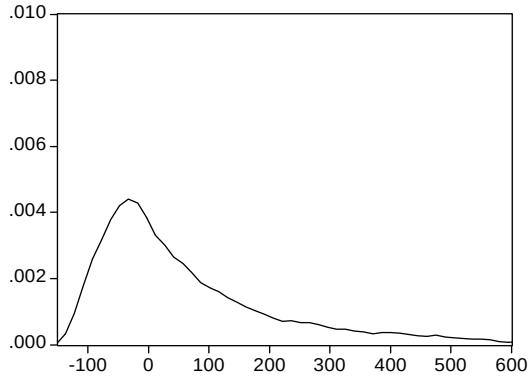


Figure 13: distribution of put_error_1999 (%)

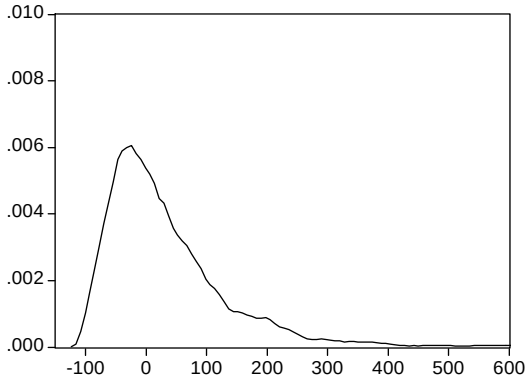


Figure 14: distribution of put_error_2000 (%)

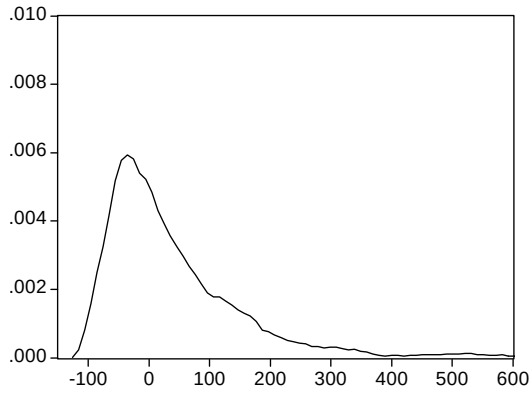


Figure 15: distribution of put_error_2001 (%)

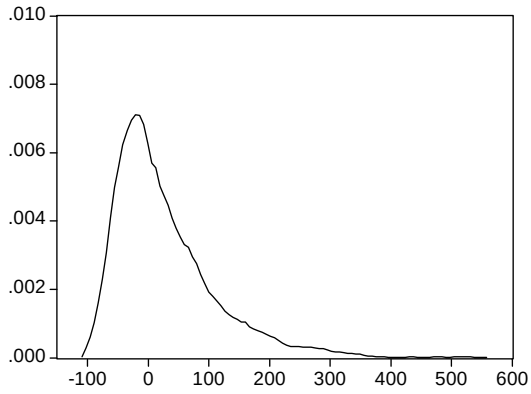


Figure 16: distribution of put_error_2002 (%)

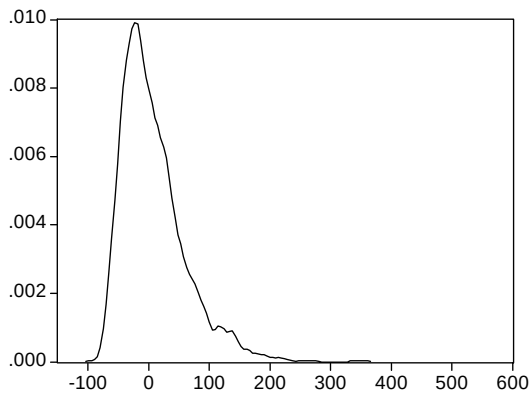
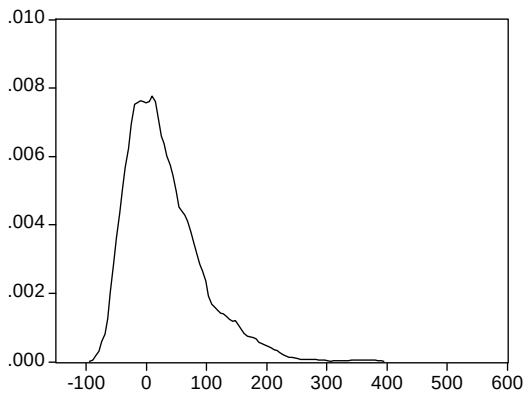


Figure 17: distribution of put_error_2003 (%)



4.5. Synthesis and conclusion

We saw that arbitrage assumptions are not satisfied when dealing with real estates; the most salient of difficulties being the assumption of the existence of a riskless portfolio, and the possibility of taking advantage of the theoretical arbitrage opportunities. The reality seems to be far from the arbitrage paradigm when dealing with that kind of assets. Consequently the validity of the valuations based on a PDE is not obvious; in section 4.4 we saw that the error could easily be very significant for a put option. A rough estimation of the options prices is maybe as good as the very complicated and not fully convincing method in a costs/benefits view.

This paper could be seen as a part of a more general topic, concerning the validity of real options. Options are a very powerful tool for the stock and the interest rates markets. This success has led numerous attempts to expand the concepts to other asset classes, however the transposition is not always straightforward and without risks. All the implicit assumptions must be re-examined rigorously before being applied to a less "pure" financial field. Because a blind application of the concepts could result in biased pricings, the extension must be conducted carefully and step by step. Examples of these problems can be found in Lautier (2002a and 2002b) or Philippe (2004).

Childs et al. (2001-2002a-2002b-2004) have provided interesting concepts, well adapted to the real estate specificities, and allowing to handle the uncertainty on H(0) (they employ the term "noisy asset", coming from real options theory in this situation). The tools they have developed can be used in the management of research and development projects, for the exploitation of corporate assets as mines, and for real estates. A common point between these fields is the ignorance of the exact prices unless an irreversible action is undertaken (beginning of the production or the exploitation, or sale completion). The principles of the modelling are as follows: The real price process x(t) is unknown for investors because of a noise process y(t). The only thing that can be observed is a noisy estimation z(t) of the true value x(t) , the relation can be for instance $z(t) = x(t) + y(t)$

or z(t) = x(t) y(t). At any time t, the available information is represented by a filtration {I(t)}, and investors estimate x(t) given this filtration, by m(t) = E [x(t) | I(t)] . This value m(t) is an appraisal value conditional on all the relevant information (for instance sale prices for similar real estate), and is named the time-filtered asset value. Subsequently using Lipster and Shiryayev differential equations, the dynamic of m(t) can be expressed with the parameters present in the dynamics of x(t) and y(t). Once this is done Childs et al price an option on a noisy asset x(t) with a version of Black-Scholes formula but using m(t) as underlying. Unfortunately, the same concerns discussed above for the PDE could be raised for this pricing methodology. m(t) is not a tradable asset contrary to the stock in the Black-Scholes world. Moreover, at maturity the payoff is not m(T), but x(T), making the construction of a riskless portfolio difficult, which puts the validity of the pricing formula under question. Similarly, for t < T, m(t) is an optimal appraisal price, it is not the real price needed to build an actual arbitrage portfolio. In Childs et al, a pricing formula is not really the aim, the analysis rather stresses the importance of information, comparing the costs and the benefits. Indeed, as the acquisition of information is costly, it is necessary to estimate its usefulness. This framework is more intended to be an analysis of economical choices rather than a financial pricing, and this shifting is not really surprising. Investors will decide exogenously for the premium associated to the risk generated by the noise around the valuation. In other words, if they think that it can be sufficiently valuable to acquire information they will proceed for it. But this decision relies on their appetite for risk, and on their personal valuation of the costs associated with this uncertainty. It is an economic choice. The problem of valuation of real options can be dealt with financial theory, but the issue of the usefulness of information cannot be ignored. Fundamentally real estate markets are incomplete, in the sense that all possible portfolio positions are not attainable. For example if someone owns a specific house with a specific service flow, a hypothetical trader wishing to arbitrage on a prepayment option could not include exactly this particular house in his portfolio, because of its uniqueness. In addition

to this physical incompleteness, there exists an informational incompleteness coming from the appraisal problem. How could be constituted a portfolio relying on H(t) if this value in unknown? Theory says that in incomplete markets it does not exist a single no-arbitrage price but only an interval of prices compatible with this no-arbitrage assumption. As far as it concerns the noise issue things can be understood heuristically. If there exist noise on H it would be surprising to obtain an exact formula for V(t, r, H); dispersion in figure 8 for H(0) leads to a dispersion in figure 9 for P(0). This price interval represents the economical freedom associated with the noise risk and the possibility for each agent to price this uncertainty according to his own preferences.

The specific features of real estates make the pricing of contingent claims uneasy. Properties are strongly segmented and this lack of a global and uniform market prevents a purely financial approach. Real estate is not a partitive asset, in the sense that you cannot buy "some" real estate as you can buy some sugar. In a financial market one can buy some IBM (stock), some oil, or some treasury bond, nearly perfectly assimilated with another stock, another barrel or another bond. The uniqueness of houses hinders this financial view and, in a way, real estate markets do not exist; we can only speak of multiple local markets each one with its particularities. This situation is very far from the assumptions used in the financial models and the lack of a product globally traded, linked to the real estates, is probably one of the major obstacles to a purely financial approach. A solution could be in a reversal of this situation. Instead of viewing the real estate as the underlying asset, the primary asset could be a property index or a derivative on this index. If such a market becomes sufficiently well developed it could provide a price for "some" real estate, making easier a financial pricing for contingent claims.

5. Conclusion

Dans ce dernier chapitre trois voies d'approfondissement ont été ouvertes. La notion d'indice informationnel a d'abord été définie en toute généralité et en adoptant une démarche de quantification explicite de l'information. Définir l'information, ou la représentativité d'un bien par rapport à un stock, semble en effet être un passage incontournable pour pouvoir travailler avec le concept de prix dans les situations de marchés hétérogènes et illiquides. Dans une optique plus abstraite, l'analyse de la structure de la matrice Î a été initiée à l'aide de concepts géométriques. Ces outils rendent possible les études fines et détaillées des distributions informationnelles. Enfin, sur un plan plus directement économique, l'applicabilité de la théorie de l'arbitrage aux sous-jacents immobiliers a été discutée. Si cette approche peut apparaître, au premier abord, comme la plus naturelle pour évaluer des produits dérivés sur indices immobiliers, nous avons pu nous rendre compte que cela ne se fera pas sans difficultés en raison des particularités et des spécificités des sous-jacents.

Conclusion

Dans ce dernier chapitre, qui fera office de conclusion, nous récapitulerons tout d'abord les principaux résultats de cette thèse. Des voies d'approfondissements et des perspectives pour de futures recherches seront ensuite présentées, puis nous discuterons du rôle central de la quantification informationnelle dans les marchés hétérogènes et illiquides. Cette question est en effet apparue au cours de ce travail comme étant au cœur de la problématique indicielle. Enfin, la dernière section conclura en évoquant les points communs qui existent entre l'immobilier et le marché de l'art et elle discutera de l'applicabilité du RSI à ce champ d'investissement. La similarité des caractéristiques financières de ces deux marchés est en effet une invitation à adopter une position abstraite et générale pour rendre compte de ces deux environnements.

1. Les contributions scientifiques de cette thèse

En prenant comme point de départ l'étude du lien théorique entre les indices de prix et les indices de rendement, un nouveau formalisme maniable, interprétable, et performant a été élaboré pour le modèle des ventes répétées. Il nous a permis d'aborder, avec les mêmes concepts fondamentaux et unificateurs, différents aspects et problèmes du RSI qui, jusqu'à présent, pouvaient apparaître comme n'ayant qu'un lointain rapport. Cet exercice de réécriture formelle constitue, par ses conséquences pratiques, le principal apport théorique de ce travail. La liste ci-dessous récapitule les contributions les plus significatives qui ont découlé de cette démarche.

- Nous avons établi l'existence d'un lien fonctionnel entre les indices de prix et l'indice de ventes répétées[1].

[1] Perturbé toutefois par des coefficients aléatoires

- En combinant les concepts théoriques du modèle avec une technique de construction d'échantillons de référence, une méthodologie d'analyse de données a été développée. Elle permet d'approfondir très significativement l'étude économique des ventes répétées et de ne plus se contenter des seules valeurs indicielles. De multiples indicateurs viennent enrichir l'analyse des comportements des participants au marché, augmentant ainsi très notablement l'extraction de l'information enchâssée dans les bases de données.

- A partir d'un générateur de données synthétiques, une étude de la sensibilité de l'indice aux paramètres du modèle a été réalisée. Plusieurs phénomènes nouveaux, non répertoriés dans la littérature et dont les effets ne sont absolument pas négligeables, ont été mis en évidence. Les résultats les plus marquants concernent les effets de bords, les chocs de liquidité et leurs effets parfois paradoxaux sur la fiabilité de l'indice, ainsi que le rôle de l'asymétrie. Cette dernière caractéristique, par son impact et sa capacité explicative, s'est révélée être de première importance.

- Les concepts théoriques d'information et de bruit, introduits au niveau élémentaire des transactions, engendrent au niveau agrégé de l'indice des formules simples, intuitives et cohérentes pour :

  - la volatilité : $\mathcal{V}(R) = \sigma_G^2 \hat{I}^{-1}$
  - la réversibilité : $\hat{I}(T_2)\, R(T_2) = \hat{I}(T_1)\, R(T_1) + \hat{I}(T_2 \backslash T_1)\, R(T_2 \backslash T_1)$

  Cette dernière formule, par sa maniabilité et son interprétabilité, est supérieure à celle proposée par Clapp et Giaccotto (1999).

- En utilisant cette formule théorique de réversibilité, une méthodologie de quantification empirique des fluctuations potentielles a pu être mise en place. Il n'existait pas, jusqu'alors, dans la littérature de technique comparable permettant de se faire une idée de l'ampleur de ce phénomène indésirable. Or,

dans la perspective de l'introduction des produits dérivés sur indice immobilier, la volatilité rétroactive du RSI doit impérativement pouvoir être appréhendée.

- L'estimation d'un indice sur un échantillon constitué de plusieurs populations se révèle parfois délicate, car les caractéristiques financières du RSI ne sont pas toujours représentatives du marché dans son ensemble. Il peut arriver qu'un segment du marché domine les autres dans son expressivité vis-à-vis de l'indice. Le risque immobilier moyen est alors mal perçu par les investisseurs et cette situation peut être à l'origine d'allocations d'actifs sous-optimales.

2. Les perspectives pour de futures recherches

A partir de ces résultats, plusieurs pistes de recherche sont envisageables ; certaines d'entre elles ont d'ailleurs déjà été évoquées dans le chapitre 4.

Dans une optique économique, la mise en œuvre de la méthodologie d'analyse de données du chapitre 2 pourrait se révéler instructive pour l'étude des comportements des propriétaires sur la période 1980-2005. Comment réagissent par exemple les différents indicateurs (incitation à l'achat, incitation à la revente, longueur de la période de détention…) pendant la période de bulle immobilière et au cours du krach qui l'a suivi ? Les résultats d'une telle étude confirmeront-ils les analyses qui ont déjà été menées sur ces événements très marquants pour le marché immobilier, ou bien mettront-ils en évidence des phénomènes nouveaux en s'appuyant sur la diversité des indicateurs disponibles ?
Sur un autre plan et toujours en se basant sur des données réelles, la technique de quantification empirique de la réversibilité pourrait être testée concrètement et affinée. En particulier, il serait utile d'étudier le nombre de sources aléatoires à

prendre en compte dans les simulations pour rendre les résultats optimaux. On peut avancer l'hypothèse que ce chiffre sera très probablement lié au contexte économique.

Une autre direction possible concerne ce que l'on pourrait appeler, de manière générale, les méthodologies de correction informationnelle. Nous avons pu constater qu'il arrive parfois que certaines données soient surreprésentées ou trop expressives dans les échantillons d'estimation, par exemple dans les situations de population dominante ou les cas d'asymétrie très marquée. Dans de tels environnements, les indices que l'on obtient retranscrivent assez mal le risque immobilier moyen et ils sont souvent affectés par des erreurs d'estimation non négligeables.
Pour résoudre ce problème, la première idée qui vient à l'esprit consiste à réduire les échantillons d'estimation, en leur enlevant certaines des données en excès. Mais, jusqu'où faut-il les réduire et selon quel critère ? Car il est bien sûr évident qu'une application téméraire et directe de cette idée relèverait plus de la recette de cuisine que d'une approche économétrique rigoureuse et scientifique. Sur ce point, la quantification informationnelle des ventes répétées pourrait se révéler très utile pour fonder solidement cette technique de restriction d'échantillons. Ainsi à titre d'illustration et en se contentant uniquement de l'intuition : Si dans un modèle à deux populations, la population A est informationnellement deux fois plus importante que la population B, une réduction légitime d'échantillon pourrait avoir pour but de ramener les quantités d'information à des niveaux comparables. Au passage il faut remarquer que l'on ne raisonne pas ici en termes d'effectifs bruts, mais avec des grandeurs informationnelles. Une telle technique serait également envisageable pour les situations asymétriques, ou pour tout autre problème de surreprésentation.

Tous ces raffinements économétriques, et plus généralement toutes les recherches entreprises sur l'immobilier depuis ces vingt dernières années, témoignent d'une financiarisation de plus en plus marqué de ce secteur. L'exemple le plus frappant de ce phénomène concerne les indices. Ils sont actuellement en train de

passer du statut d'indicateurs économiques classiques à celui d'actifs négociables à part entière, des produits dérivés les utilisent déjà comme sous-jacents. Très rapidement, il faudra donc pouvoir valoriser correctement ces titres financiers, car l'existence de méthodes d'évaluation fiables et performantes est en général requise pour qu'un marché de ce type puisse devenir mature. L'étude des produits dérivés écrits sur un indice de ventes répétées constitue donc un prolongement naturel pour ce travail.

Cependant, comme cela a déjà été souligné dans le chapitre 4 (paragraphe 4) avec l'exemple de la théorie de l'arbitrage appliquée aux options implicites enchâssées dans les prêts immobiliers, la difficulté et la complexité de cette entreprise ne doivent pas être sous-estimées. Les spécificités de l'actif immobilier sont en effet incontournables et les hypothèses de la théorie financière, formulées pour des marchés de titres, ne sont pas automatiquement transposables aux marchés immobiliers. Or nous savons bien que dans la démarche scientifique une modélisation raisonnable ne peut pas se fonder sur des caractéristiques souhaitables pour l'objet, elle doit s'appuyer sur les caractéristiques réelles de celui-ci. L'impasse à laquelle nous avons été confrontés dans ce chapitre 4, lors de l'étude critique des modèles d'arbitrage classiques appliqués aux « mortgages » est en fait probablement symptomatique d'une rupture épistémologique. Les marchés financiers modernes disposant de bonnes propriétés (liquidité, fongibilité, profondeur, cotation…) se prêtent assez bien à l'exercice de la quantification, par contre pour les marchés hétérogènes et illiquides la démarche, sans être impossible, est certainement un peu moins immédiate. En d'autres termes, la finance des actions ce n'est pas la finance de l'immobilier, et vice-versa.

Si le paragraphe 4 du chapitre 4 a pointé les difficultés des raisonnements par arbitrage lorsque le sous-jacent est un actif immobilier direct et, puisqu'un indice présente de meilleures propriétés en termes de modélisation qu'un bien particulier, on peut alors se demander légitimement si les objections à l'emploi de cette théorie ne seraient pas levées pour l'évaluation des dérivés indiciels. La situation est en fait certainement beaucoup plus favorable ici. Mais, elle n'autorise pas pour autant à

ignorer les spécificités immobilières et la prudence s'impose encore. Considérons en effet l'hypothèse faite sur la quasi-totalité des processus stochastiques en finance et qui est fondamentalement liée au concept d'efficience des marchés : le comportement markovien[2]. Peut-on modéliser la dynamique d'un indice de ventes répétées avec un processus de ce type ? La réponse est probablement négative, en raison du phénomène de réversibilité. On peut s'en rendre compte intuitivement en réécrivant la formule associée sous la forme : $\hat{I}(T_2)R(T_2) - \hat{I}(T_1)R(T_1) = \hat{I}(T_2 \backslash T_1)\ R(T_2 \backslash T_1)$. La partie gauche de cette égalité mesure un accroissement entre le présent T_1 et le futur T_2. Si l'hypothèse de Markov est satisfaite, cette variation ne doit pas dépendre des dates antérieures à T_1. Or, on sait que le terme de droite $\hat{I}(T_2 \backslash T_1)R(T_2 \backslash T_1)$, qui est associé aux nouvelles données arrivées lors de l'extension de l'horizon d'estimation de T_1 à T_2, ne se contente pas d'apporter de l'information sur l'intervalle $[T_1,T_2]$; il renseigne en fait sur l'intégralité de l'intervalle $[0,T_2]$. Le RSI ne peut donc pas avoir un comportement markovien. Par conséquent, l'utilisation des mouvements browniens (processus markoviens) pour décrire la dynamique d'un indice de ventes répétées n'est pas une évidence et cela demanderait de mures réflexions[3]. L'enjeu de cette modélisation, c'est-à-dire la possibilité d'évaluer sur des bases solides les dérivés, est cependant majeur pour l'industrie de la finance immobilière.

On pointera enfin une autre piste susceptible de faire l'objet de futurs approfondissements. Elle ne concerne pas directement des problématiques empiriques d'estimation indicielle, des questions économiques et comportementales ou des enjeux industriels comme les précédentes, mais elle n'est pas pour autant

[2] On dit qu'un processus est markovien si son futur ne dépend du passé que par l'intermédiaire de l'état présent. En d'autres termes, le trajet qui a été suivi par un processus pour arriver à un niveau X_s à la date s n'influe pas sur les probabilités de réalisation de son futur X_t ($t > s$). Financièrement, cette hypothèse mathématique est une des formulations possibles pour le concept d'efficience des marchés. La valeur présente X_s contient toute l'information passée, il n'est donc pas nécessaire d'aller rechercher dans son histoire des éléments permettant de déterminer un meilleur niveau pour le processus à la date s. Le marché a en effet déjà agrégé toute l'information pertinente en fixant le prix à X_s.

[3] Peut-être faudrait-il décrire le risque de réversibilité par un processus stochastique à part entière, et l'intégrer ensuite à la modélisation de la dynamique de l'indice de ventes répétées.

dénuée de légitimité et d'intérêts. Le paragraphe 3 du chapitre 4 a tenté de poser quelques bases pour une étude formelle des distributions informationnelles à l'aide d'outils géométriques. Cette démarche mathématique, dont les applications concrètes ne sont pas immédiatement discernables, s'inscrit de manière tout à fait cohérente dans le prolongement de cette thèse, dont l'un des apports principaux a consisté à mettre en évidence la structure informationnelle implicite et fondamentale du RSI. Le rôle primordial de la mesure de l'information dans les indices immobiliers laisse donc à penser qu'une étude un peu abstraite de la structure des matrices d'information, et des volumes informationnels associés, ne serait pas vaine. Des répercussions opérationnelles devraient pouvoir émerger d'une approche formaliste de cette nature.

3. Le rôle central de la quantification informationnelle dans les marchés hétérogènes

Comme en témoigne, entre autres choses, la dernière piste de recherche présentée ci-dessus, le concept d'information s'avère être le concept central pour l'étude des marchés hétérogènes. La mise en lumière de la structure sous-jacente à l'indice de ventes répétées, dans sa version classique, a permis sa généralisation ; le concept global d'indice informationnel a ainsi été introduit dans le paragraphe 2 du chapitre 4. Cette nouvelle approche a l'avantage de rendre explicite la quantification de l'information alors que l'estimation traditionnelle par moindres carrés la dissimule. Pour être plus exact, comme tout calcul d'indice pour un marché hétérogène ne peut pas se faire sans répondre d'une manière ou d'une autre à cette question[4], la méthode

[4] Pour la méthode des ventes répétées cette affirmation est étayée par le travail développé dans cette thèse. Pour les autres types d'indices, elle reste d'ordre intuitif. Il pourrait être intéressant d'essayer de reformuler la méthodologie hédonique traditionnelle, en s'efforçant de dégager ce qui pourrait s'apparenter à une structure informationnelle, afin de confirmer ou d'infirmer cette hypothèse générale.

classique fournit en fait indirectement une réponse automatique, catégorique et tranchée à un problème qu'elle ne se pose pas. Ignorer cette problématique informationnelle revient alors à renoncer à l'usage d'une modélisation souple, maniable et cohérente.

Si l'on décide, à l'opposé, d'adopter une stratégie de quantification explicite, le point central est alors de déterminer le niveau informationnel des transactions. Comment faut-il décider du degré de représentativité d'un prix de vente ou d'un prix d'achat par rapport à cette notion abstraite de « prix de l'Immobilier »? Ou, plus laconiquement, combien y a-t-il d'Immobilier dans une maison ou un appartement donné ? Comme mentionné antérieurement dans le chapitre 4, certaines variantes du RSI, déjà répertoriées par la littérature indicielle, reviennent en fait à adopter des définitions alternatives pour l'information. Le champ des possibles est cependant très loin d'avoir été épuisé et beaucoup d'autres définitions informationnelles pourraient être envisagées.

« Le » prix de l'immobilier est, comme on nous l'avons vu dès l'introduction générale, une notion mal définie et sans doute indéfinissable. Cette reformulation de la problématique, qui nous amène à substituer à l'ambition catégorique et unitaire du chiffre unique ce nouveau paradigme informationnel plus souple, découle en fait directement et fondamentalement de l'hétérogénéité des actifs immobiliers. On ne se trouve plus dorénavant dans une situation où une seule et « vraie » courbe rend compte du marché immobilier dans son ensemble. Un ruban indiciel vient remplacer la courbe unique et, selon la définition que l'on choisira pour l'information, l'indice estimé fluctuera à l'intérieur de cette bande (figure 3 du chapitre 4).

4. Quel rapport y a-t-il entre un pavillon de banlieue et Delacroix, ou entre une chambre de bonne et une commode Louis XV ?

Bien que cette recherche sur l'indice de ventes répétées ait été menée en choisissant l'immobilier comme support, elle est cependant largement pertinente pour tout type de marché possédant des caractéristiques financières semblables. Dans une certaine mesure, la réflexion développée au cours de ce travail pourra donc être considérée comme une étude générale de la finance des marchés hétérogènes et illiquides. Avec l'immobilier nous rangerons dans cette catégorie le marché de l'art, dont les propriétés en font l'autre grand exemple de cette classe. Pour étayer un peu plus la légitimité de ce rapprochement, nous allons pointer dans ce dernier paragraphe les parallèles qui existent entre ces deux champs d'investissements[5] et nous examinerons l'applicabilité et l'intérêt du RSI pour ce secteur.

Il faut tout d'abord remarquer qu'en raison de la diversité des biens artistiques, la notion de « prix de l'art » est aussi ambiguë que celle de « prix de l'immobilier ». Dans l'introduction de son ouvrage consacré aux méthodes de gestion appliquées au marché de l'art, Mahé de Boislandelle (2005) choisit cependant de mettre l'accent sur l'action et la nécessité de la décision. Car, comme il a déjà été signalé au tout début de cette thèse dans le cas de l'immobilier, ce n'est pas parce que les notions sont un peu floues théoriquement qu'il faut renoncer à gérer et à optimiser. Bien au contraire, ces situations sont en effet souvent des invitations à faire preuve d'inventivité et de créativité afin de réussir à élaborer de nouvelles méthodes, mieux adaptées aux spécificités de ces champs complexes.

En ce qui concerne les praticiens, les enjeux et les contextes financiers de ces deux marchés présentent également des similarités. Ainsi, à l'image d'IPD (Invesment Property Database) pour l'immobilier, la création de la société Artprice

[5] L'immobilier et le marché de l'art ne relèvent pas uniquement de problématiques d'investissements. Ils comportent aussi, tous les deux, une dimension consommation.

(www.artprice.com) témoigne par exemple d'un besoin croissant d'informations fiables et standardisées de la part des différents acteurs du marché de l'art (maisons de ventes aux enchères, marchands, collectionneurs, galeries, assureurs, banquiers, presse, administrations…).
Il faut d'ailleurs souligner l'importance de ces deux sources informationnelles pour les problématiques relevant de la gestion de patrimoine. Pour les gérants les indices et les méthodologies d'estimation hédonique constituent en effet des outils précieux pour appréhender la valeur d'un patrimoine et agir en conséquence. Dans une perspective plus académique, l'intégration explicite par le truchement des indices de l'actif immobilier, ainsi que des objets d'art, dans le portefeuille des ménages est également très intéressante pour analyser la détention des titres financiers classiques par ces acteurs économiques. Car ne pas prendre en compte ces deux types d'investissement, immobilier et artistique, qui représentent souvent une part très importante du patrimoine privé, fait encourir le risque de mal évaluer le niveau d'incertitude auquel doit faire face le particulier.

Sur le plan de la technique indicielle stricto sensu, on pourra se référer à Morieux (1980) pour avoir un exemple d'emploi de la méthodologie hédonique et à Goetzmann (1993) pour les ventes répétées. Le premier étudie l'évolution du prix des commodes Louis XV entre 1949 et 1970, le second le prix des peintures sur la période 1715-1986. Nous retrouvons donc ici les mêmes techniques économétriques que pour l'immobilier. Le RSI semble, par certains côtés, assez adapté au marché de l'art. L'objection sur la qualité non constante des biens immobiliers, qui est parfois évoquée comme un biais méthodologique, est beaucoup moins légitime dans le cas d'une œuvre d'art puisqu'elle n'est que très rarement modifiée, c'est-à-dire essentiellement restaurée, ou altérée. Il faudra bien sûr éviter d'appliquer aveuglement les résultats obtenus dans un contexte immobilier à un environnement relevant du marché de l'art. Les spécificités de ce domaine ne doivent en effet pas être ignorées en prétextant abusivement d'une structure générale et commune car, s'il est vrai que beaucoup d'éléments sont semblables, il n'y a pas pour autant identité.

Ainsi, la segmentation en sous-marchés est sans doute plus marquée (un artiste = un marché) et les phénomènes d'engouement ou de désuétude probablement plus amples et plus fréquents. La souplesse et la fonctionnalité de l'approche informationnelle abstraite, développée dans le chapitre 4, devraient toutefois pouvoir permettre d'élaborer des méthodologies de ventes répétées adaptées au marché de l'art. Les deux problèmes évoqués ci-dessus pourraient par exemple se gérer en adaptant les quantités d'information associées à chaque bien. Un artiste passé de mode serait ainsi moins informationnel pour l'indice qu'une valeur sure comme Picasso ou Van Gogh.

Comme nous pouvons le constater, les points communs et les convergences entre marché immobilier et marché de l'art ne sont pas mineurs. Dans une certaine mesure, les étudier en parallèle est donc fondé et légitime. Mais plus qu'une étude en parallèle, il faudrait peut-être envisager une approche conceptuelle unifiante pour ces situations d'hétérogénéité et d'illiquidité marquées…

Annexes

Annexe 1 : Extension des concepts à l'échantillon considéré dans son ensemble

$N = \sum_{i<j} n_{ij}$ sera la généralisation de $n^t = \sum_{i \le t < j} n_{i,j}$

$B_i^t = L_{i,t+1} + L_{i,t+2} + \ldots + L_{i,T}$ et $S_j^t = L_{0,j} + L_{1,j} + \ldots + L_{t,j}$ engendreront :

$B_i = L_{i,i+1} + L_{i,i+2} + \ldots + L_{i,T}$ et $S_j = L_{0,j} + L_{1,j} + \ldots + L_{j-1,j}$

$B^t = B_0^t + \ldots + B_t^t$ et $S^t = S_{t+1}^t + \ldots + S_T^t$ se réécriront :

$B = B_0 + \ldots + B_{T-1}$ et $S = S_1 + \ldots + S_T$

L'équivalent de I^t est $I = \sum_{i<j} L_{ij}$; la relation $B^t = S^t = I^t$ sera transformée en $B = S = I$.

Les concepts B_i^t, S_j^t, B^t, S^t sont des concepts informationnels, par conséquent on pourra aussi définir leur équivalents réels b_i^t, s_j^t, b^t, s^t en transposant naturellement les formules.

On pourra aussi généraliser $\tau^t = \left(I^t / n^t \right)^{-1}$ par $\tau = \left(I / N \right)^{-1}$ et $F^t = I^t / n^t$ par $F = I/N$. F sera toujours une moyenne arithmétique et τ une moyenne harmonique.

$H_p(t)$, la moyenne géométrique des $h_{0,\ldots,} h_t$ avec les poids B_i^t, se transformera en H_p, moyenne géométrique des $h_{0,\ldots,} h_{T-1}$ pondérée par les B_i. Il en sera de même pour $H_f(t)$ et H_f.

On se reportera au paragraphe 6 du chapitre 1 pour un résumé exhaustif, et en toute généralité, des différentes notations du modèle.

Annexe 2 : Reformulation de la première somme dans $\partial\Phi(R) / \partial r_t$

La somme $\sum_{i \le t < j} L_{i,j} (r_i + \ldots + r_{j-1})$ peut être étudiée grâce au tableau suivant :

Date de revente

Date d'achat : i \ j	t + 1	t + 2	…	T – 1	T
0	$(r_0+\ldots+r_t)$ $L_{0,\,t+1}$	$(r_0+\ldots+r_{t+1})$ $L_{0,\,t+2}$		$(r_0+\ldots+r_{T-2})$ $L_{0,\,T-1}$	$(r_0+\ldots+r_{T-1})$ $L_{0,\,T}$
1	$(r_1+\ldots+r_t)$ $L_{1,\,t+1}$	$(r_1+\ldots+r_{t+1})$ $L_{1,\,t+2}$		$(r_1+\ldots+r_{T-2})$ $L_{1,\,T-1}$	$(r_1+\ldots+r_{T-1})$ $L_{1,\,T}$
2	$(r_2+\ldots+r_t)$ $L_{2,\,t+1}$	$(r_2+\ldots+r_{t+1})$ $L_{2,\,t+2}$		$(r_2+\ldots+r_{T-2})$ $L_{2,\,T-1}$	$(r_2+\ldots+r_{T-1})$ $L_{2,\,T}$
⋮					
T – 1	$(r_{t-1}+r_t)$ $L_{t-1,\,t+1}$	$(r_{t-1}+\ldots+r_{t+1})$ $L_{t-1,\,t+2}$		$(r_{t-1}+\ldots+r_{T-2})$ $L_{t-1,\,T-1}$	$(r_{t-1}+\ldots+r_{T-1})$ $L_{t-1,\,T}$
T	r_t $L_{t,\,t+1}$	(r_t+r_{t+1}) $L_{t,\,t+2}$		$(r_t+\ldots+r_{T-2})$ $L_{t,\,T-1}$	$(r_t+\ldots+r_{T-1})$ $L_{t,\,T}$

Les "r_0" sont fournis par la première ligne, les "r_1" par la première et la seconde, …, les "r_t" par les t+1 lignes. On obtient les "r_{T-1}" de la dernière colonne, les "r_{T-2}" des deux dernières colonnes, …, les "r_{t+1}" de toutes les colonnes sauf la première.

Ainsi pour $t' \le t$, le nombre de $r_{t'}$ est :

$(L_{0,\,t+1} + L_{0,\,t+2} + \ldots + L_{0,\,T}) + (L_{1,\,t+1} + L_{1,\,t+2} + \ldots + L_{1,\,T}) + \ldots + (L_{t',\,t+1} + L_{t',\,t+2} + \ldots + L_{t',\,T})$

(Ligne 0) (Ligne 1) (Ligne t')

A savoir: $B_0^t + B_1^t + \ldots + B_{t'}^t$

Et pour $t' > t$, le nombre de $r_{t'}$ est :

$(L_{0,\,T} + L_{1,T} + \ldots + L_{t,\,T}) + (L_{0,\,T-1} + L_{1,T-1} + \ldots + L_{t,\,T-1}) + \ldots + (L_{0,\,t'+1} + L_{1,t'+1} + \ldots + L_{t,\,t'+1})$

(Colonne T) (Colonne T-1) (Colonne t'+1)

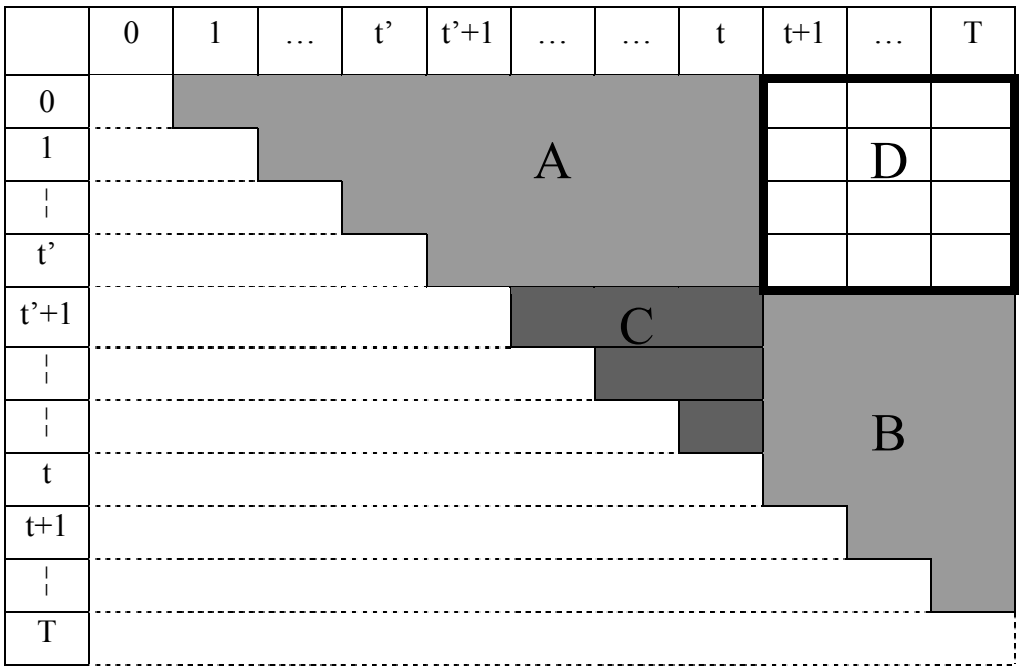
C'est-à-dire : $S^t_T + S^t_{T-1} + \ldots + S^t_{t'+1}$

Annexe 3 : Calcul de $I^{[t', t+1]}$ pour l'échantillon benchmark

Dans la figure 1, $I^{[t', t+1]}$ correspond à la somme des $L_{i, j}$ contenus dans le rectangle D. On sait que la somme des $L_{i,j}$ pour l'aire totale (A+B+C+D) est égale à I et que son expression est fournie par la formule (25) du chapitre 1. On peut, de plus, remarquer que l'aire D peut se réécrire :

$$D = I - (A + C) - (B + C) + C$$

Figure 1 : information pertinente pour [t', t+1], t' < t



Les aires A + C, B + C et C étant triangulaires et du même type que l'aire totale, on peut utiliser la formule (25) du chapitre 1, en adaptant le paramètre de taille T, pour obtenir la somme de leurs éléments. On profite ici de la forte régularité des $L_{i, j} = K' \alpha^{j - i} / (j - i)$ pour le benchmark. La seule chose qui importe dans cette distribution est en fait la distance à la diagonale[1], mesurée par j – i.

[1] D'une certaine manière, on pourrait qualifier cette distribution d'homothétique.

Les tailles sont de : T pour A + B + C + D t pour A + C
T – t' – 1 pour B + C t – t' – 1 pour C

La somme des éléments de D s'obtient alors en écrivant :

$$\begin{aligned} I^{[t',t+1]} = \; & K' ((T+1)\, u_T - (\alpha / (1-\alpha)) * (1 - \alpha^{T})) \\ & - K' ((t+1)\, u_t - (\alpha / (1-\alpha)) * (1 - \alpha^{t})) \\ & - K' ((T-t')\, u_{T-t'-1} - (\alpha / (1-\alpha)) * (1 - \alpha^{T-t'-1})) \\ & + K' ((t-t')\, u_{t-t'-1} - (\alpha / (1-\alpha)) * (1 - \alpha^{t-t'-1})) \end{aligned}$$

D'où : $$I^{[t',t+1]} = K'[(u_T - u_t) + t'(u_{T-t'-1} - u_{t-t'-1}) - t(u_t - u_{t-t'-1}) + T(u_T - u_{T-t'-1})] - K [- \alpha^{T} + \alpha^{t} + \alpha^{T-t'-1} - \alpha^{t-t'-1}]$$

Introduisons la notation suivante pour alléger l'écriture :

$$U(m,n) = u_n - u_m = \alpha^{m+1} / (m+1) + \ldots + \alpha^{n} / n \qquad \text{pour } m \leq n$$

On a alors :

$$I^{[t',t+1]} = K'[U(t,T) - (tU(t-t'-1,t) - t'U(t-t'-1,T-t'-1)) + TU(T-t'-1,T)] + K \alpha^{t-t'-1} (1 - \alpha^{t'+1}) (1 - \alpha^{T-t})$$

La première partie étant encore peu maniable, on introduira une nouvelle notation en définissant une fonction $\mathcal{U}$ à trois variables (t',t, T). Pour $0 \leq t' < t < T$ on pose :

$$\mathcal{U}(t', t, T) = U(t,T) - (tU(t-t'-1,t) - t'U(t-t'-1, T-t'-1)) + TU(T-t'-1,T)$$

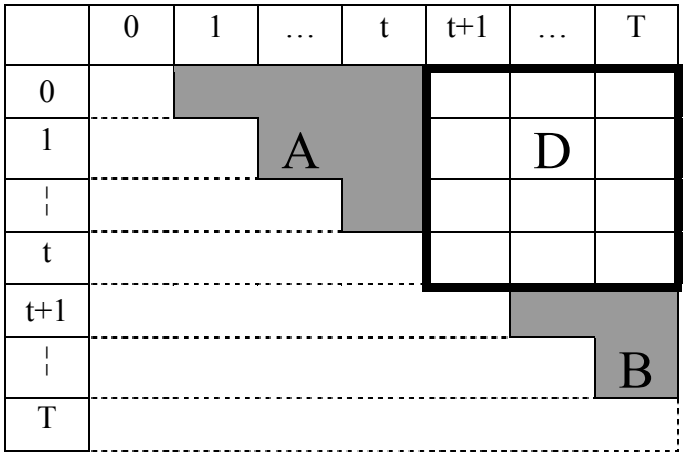
Il vient : $$I^{[t',t+1]} = K' \mathcal{U}(t', t, T) + K \alpha^{t-t'-1} (1 - \alpha^{t'+1}) (1 - \alpha^{T-t})$$

Annexe 4 : Calcul de I^t pour l'échantillon benchmark

Les formules de l'annexe 3 peuvent être appliquées si t' = t – 1, en utilisant la convention $u_0 = 0$. Géométriquement, ce cas signifie simplement que l'aire C n'existe pas et que l'on a juste : D = I – A – B.

Malheureusement, si t' = t, c'est-à-dire si l'on cherche $I^{[t, t+1]} = I^t$, la démonstration précédente ne s'applique plus. La figure 2 représente les aires en jeu dans ce cas. Non seulement l'aire C est absente, mais de plus les aires A et B sont séparées par la $(t+1)^{eme}$ colonne, contrairement à ce qui se passe pour la figure 1. Le calcul doit donc être conduit séparément.

Figure 2 : information pertinente pour [t, t+1]



Les tailles sont : T pour I t pour A T – t – 1 pour B

Le calcul de la somme des éléments de D s'effectue en remarquant que D = I - A - B

D'où :
$$I^t = K' \left((T+1)\, u_T - (\alpha / (1-\alpha)) * (1 - \alpha^T) \right)$$
$$- K' \left((t+1)\, u_t - (\alpha / (1-\alpha)) * (1 - \alpha^t) \right)$$
$$- K' \left((T-t)\, u_{T-t-1} - (\alpha / (1-\alpha)) * (1 - \alpha^{T-t-1}) \right)$$

et : $$I^t = K'[(u_T - u_t) - t\,(u_t - u_{T-t-1}) + T\,(u_T - u_{T-t-1})] - K\,[-\alpha^T + \alpha^t + \alpha^{T-t-1} - 1\,]$$

Si l'on souhaite utiliser les mêmes notations U(m,n) et $\mathcal{U}(t', t, T)$ que ci-dessus, il faut être prudent vis-à-vis du terme $(u_t - u_{T-t-1})$ car son signe varie. La solution choisie ici consiste à introduire dans la formule la valeur u_{-1} en écrivant :

$$u_t - u_{T-t-1} = u_t - u_{-1} + u_{-1} - u_{T-t-1} = (u_t - u_{-1}) - (u_{T-t-1} - u_{-1})$$

Pour être parfaitement rigoureux, il faudrait préciser la valeur de u_{-1} . Toutefois, comme il s'agit uniquement d'un artefact de calcul, ce choix n'aura en fait aucune influence sur le résultat final. On pourra donc choisir un réel quelconque. Dans un souci de simplicité, on prendra $u_{-1} = 0$.

On a alors :

$$I^t = K'[U(t,T) - (\,t\,U(-1,t) - t\,U(-1,T-t-1)\,) + TU(T-t-1,T)] + (K/\alpha)\,[\,(1-\alpha^{t+1})\,(1-\alpha^{T-t}) - (1-\alpha)\,]$$

L'introduction de u_{-1} permet maintenant de généraliser directement la définition de la fonction $\mathcal{U}(t', t, T)$ pour $t' = t$. Le résultat final s'écrit donc :

$$I^t = K'\,\mathcal{U}(t, t, T) + (K/\alpha)\,[\,(1-\alpha^{t+1})\,(1-\alpha^{T-t}) - (1-\alpha)\,]$$

Annexe 5 : Etude de la fonction F

Dans les calculs suivants la variable α ne sera pas systématiquement indiquée afin d'éviter des écritures trop chargées.

Sens de variations de F

$$F(\alpha) = I(\alpha) / N(\alpha) \text{ et } F' = (I'N - IN') / N^2$$

$$\pi(\alpha) = (\alpha / T (1 - \alpha)) * (1 - \alpha^T)$$

$$\begin{aligned}\pi'(\alpha) &= [1 - \alpha^T - T \alpha^T (1 - \alpha))] / (T(1 - \alpha)^2) \\ &= [T (1 - \alpha) \pi(\alpha) / \alpha - T \alpha^T (1 - \alpha)] / (T(1 - \alpha)^2) \\ &= [\pi(\alpha) / \alpha - \alpha^T] / (1 - \alpha)\end{aligned}$$

En utilisant le tableau 5 du chapitre 1, et en sommant sur les diagonales on a :

$$N(\alpha) = K' \sum_{i = 1,\ldots,T} \sum_{j = 1,\ldots,i} \alpha^j$$

$$N(\alpha) = K T (1 - \pi(\alpha)) \qquad \text{formule (22') du chapitre 1}$$

$$\begin{aligned}N'(\alpha) &= - K T \pi'(\alpha) = - K T [\pi(\alpha) - \alpha^{T+1}] / (\alpha (1 - \alpha)) \\ &= [N - KT (1 - \alpha^{T+1})] / (\alpha (1 - \alpha))\end{aligned}$$

Avec le tableau 6 du chapitre 1, en sommant à nouveau sur les diagonales :

$$I(\alpha) = K (1/\alpha - 1) \sum_{i = 1,\ldots,T} \sum_{j = 1,\ldots,i} \alpha^j / j$$

$$\begin{aligned}I'(\alpha) &= K (1/\alpha - 1) \sum_{i = 1,\ldots,T} \sum_{j = 1,\ldots,i} \alpha^{j-1} - (K/\alpha^2) \sum_{i = 1,\ldots,T} \sum_{j = 1,\ldots,i} \alpha^j / j \\ &= ((1 - \alpha)N - I) / (\alpha (1 - \alpha))\end{aligned}$$

$$
\begin{aligned}
F'(\alpha) &= [\,(1-\alpha)N^2 - IN + NI - KTI\,(1-\alpha^{T+1})\,] / [\,N^2\,\alpha\,(1-\alpha)] \\
&= [\,(1-\alpha)N - KTF\,(1-\alpha^{T+1})\,] / [N\alpha\,(1-\alpha)] \\
&= KT\,[\,(1-\alpha)(1-\pi) - F\,(1-\alpha^{T+1})\,] / [N\alpha\,(1-\alpha)] \\
&= [\,(1-\alpha)(1-\pi) - F\,(1-\alpha^{T+1})\,] * [KT / (N\alpha\,(1-\alpha))] \\
&= [\,1-\alpha - \alpha\, d(T)/T - F\, d(T+1)\,] * [KT / (N\alpha\,(1-\alpha))]
\end{aligned}
$$

Le deuxième crochet étant toujours positif, le signe de F'(α) est déterminé par le premier crochet. Une vérification numérique[2] effectuée sous Excel indique que, pour T variant dans [0, 1000] et α parcourant]0;1[, ce terme est toujours négatif. La dérivée F'(α) est donc négative et F décroît sur]0 ; 1[.

La limite de F en 0^+ peut être calculée grâce à la formule (30) du chapitre 1 :

$$F = ((1-\alpha)/(1-\pi)) * [(1 + 1/T)(u_T/\alpha) - \pi/\alpha)]$$

$$\pi/\alpha = (1-\alpha^T)/T(1-\alpha) \qquad \text{et} \qquad u_T/\alpha = 1 + \alpha/2 + \alpha^2/3 + \ldots + \alpha^{T-1}/T$$

Ainsi : $\lim_{\alpha \to 0+} F(\alpha) = (1+1/T) - 1/T = 1$

La limite en 1^- est un peu moins immédiate à établir :

On a $F = ((1-\alpha)/\alpha(1-\pi)) * [(1 + 1/T)\,u_T - \pi]$ et $\pi = (\alpha + \alpha^2 + \ldots + \alpha^T)/T$

$$
\begin{aligned}
\text{Or : } (1-\alpha)/\alpha(1-\pi) &= T(1-\alpha)/[\alpha\,(T - \alpha - \alpha^2 - \ldots - \alpha^T)] \\
&= T(1-\alpha)/[\alpha\,((1-\alpha) + (1-\alpha^2) + \ldots + (1-\alpha^T))]
\end{aligned}
$$

[2] Les calculs sont menés pour des valeurs allant de α = 0.01 à α = 0.99, avec un pas d'incrémentation de 0.01

$$= T / [\alpha (1 + (1-\alpha^2)/(1-\alpha) + \ldots + (1-\alpha^T)/(1-\alpha))]$$
$$= T / [\alpha (1 + (1+\alpha) + (1+ \alpha + \alpha^2) + \ldots$$
$$+ (1 + \alpha + \ldots + \alpha^{T-1}))]$$

Donc : $\lim_{\alpha \to 1^-} (1 - \alpha) / \alpha (1 - \pi) = 2 / (1 + T)$

et : $\lim_{\alpha \to 1^-} F(\alpha) = 2 [(1 + 1 / T) u_T(1) - 1] / (1 + T)$

où $u_T(1) = 1 + 1/2 + 1/3 + \ldots + 1/T$

Annexe 6 : Le problème de minimisation

Le problème d'optimisation général s'écrit :

$$\text{Min}_{R}\ [\sum_{i<j}\sum_{k'}\{\ln(p_{k',j}/p_{k',i}) - (r_i + \ldots + r_{j-1})\}^2 / \{\sigma_G^2(j-i) + 2\sigma_N^2\}]$$

Après factorisation et simplification par σ_G^2, il devient :

$$\text{Min}_{R}\ [\sum_{i<j}\sum_{k'}\{\ln(p_{k',j}/p_{k',i}) - (r_i + \ldots + r_{j-1})\}^2 / \{\Theta + (j-i)\}]$$

En développant les carrés et en ne conservant que les termes non constants, il se réécrit :

$$\text{Min}_{\{r_0,\ldots,r_{T-1}\}}\ [\Phi(R)]$$

où: $\Phi(R) = \sum_{i<j}(\Theta+(j-i))^{-1}\sum_{k'}\left[(r_i + \ldots + r_{j-1})^2 - 2\ln(p_{k',j}/p_{k',i})\,(r_i + \ldots + r_{j-1})\right]$

Comme k' varie entre 1 et $n_{i,j}$ pour chaque classe de ventes répétées (i , j), on a :

$$\Phi(R) = \sum_{i<j}(\Theta+(j-i))^{-1}\left[n_{i,j}(r_i + \ldots + r_{j-1})^2 - 2(r_i + \ldots + r_{j-1})\sum_{k'}\ln(p_{k',j}/p_{k',i})\right]$$

Et si l'on note $L_{i,j} = n_{i,j} / (\Theta + (j-i))$:

$$\Phi(R) = \sum_{i<j}L_{i,j}\left[(r_i + \ldots + r_{j-1})^2 - 2\left\{(1/n_{i,j})\sum_{k'}\ln(p_{k',j}/p_{k',i})\right\}(r_i + \ldots + r_{j-1})\right]$$

En utilisant la même technique que dans l'annexe 2, la dérivée partielle de Φ par rapport à r_t s'écrit :

$$\partial\Phi(R) / \partial r_t = 2\sum_{i\le t<j}L_{i,j}(r_i + \ldots + r_{j-1}) - 2\sum_{i\le t<j}L_{i,j}\left\{(1/n_{i,j})\sum_{k'}\ln(p_{k',j}/p_{k',i})\right\}$$

Il y a deux différences entre la situation complexe et la situation simplifiée[3]. Dans l'ancienne version, toutes les transactions étaient réalisées au niveau des valeurs indicielles h_i et h_j, qui pouvaient être interprétées comme des prix moyens. En d'autres termes, la dispersion des prix autour des valeurs de H était ignorée. Ici, la prise en compte de la variance des prix amène à remplacer la quantité $\ln(h_j/h_i)$ par $(1/n_{i,j})\sum_{k'} \ln(p_{k',j}/p_{k',i})$. Bien sûr, si $p_{k',j} = h_j$ et $p_{k',i} = h_i$, k' variant entre 1 et $n_{i,j}$, cette expression se simplifie et on retrouve $\ln(h_j/h_i)$.

La seconde différence se manifeste dans la définition des $L_{i,j}$ où le paramètre Θ apparaît désormais. L'interprétation de la distribution des $\{L_{i,j}\}$, comme l'équivalent informationnel de la distribution réelle des $\{n_{i,j}\}$, est toujours valide car $\Theta + (j - i)$ est, ici aussi, une mesure de bruit (cf. paragraphe 6.1 du chapitre 1).

[3] Pour le cas simplifié on avait : $\partial\Phi(R) / \partial r_t = 2 \sum_{i \le t < j} L_{i,j} (r_i + \ldots + r_{j-1}) - 2\sum_{i \le t < j} L_{i,j} \ln(h_j/h_i)$

Annexe 7 : Généralisation de la définition de la moyenne des taux moyens ρ_t

Avant d'analyser plus avant le problème de minimisation, il est nécessaire d'étudier les conséquences de ces modifications et, en particulier, de bien comprendre ce qui se passe pour le concept central de moyenne des taux moyens ρ_t , dont on a pu mesurer toute l'importance dans la section précédente. Dans le paragraphe 3.5 du chapitre 1, ce taux apparaissait via la seconde somme dans $\partial\Phi(R)/\partial r_t$ (formule 16). L'étude de cette expression devrait donc permettre de généraliser correctement ρ_t.

Pour les k' ventes répétées de la classe (i,j), les quantités $\ln(p_{k',j}/p_{k',i})$ correspondent simplement aux rendements continus. On peut donc définir les taux moyens continus[4] sur la période j – i par:

$$r_{k'}^{(i,j)} = \ln(p_{k',j} / p_{k',i}) / (j - i)$$

Avec cette notation, la deuxième somme de l'annexe 6 devient :

$$\sum_{i \le t < j} L_{i,j} \{(1/n_{i,j}) \sum_{k'} \ln(p_{k',j}/p_{k',i}) \} = \sum_{i \le t < j} (\Theta+(j-i))^{-1} \sum_{k'} r_{k'}^{(i,j)} (j - i)$$

$$= \sum_{i \le t < j} \sum_{k'} (j - i) / (\Theta+(j-i)) \, r_{k'}^{(i,j)}$$

Cette expression est très proche de la moyenne arithmétique des $r_{k'}^{(i,j)}$ pondérés par les coefficients $(j - i) / (\Theta + (j - i))$. Le poids $(j - i) / (\Theta + (j - i))$ peut être interprété comme la proportion du bruit provenant de la marche aléatoire gaussienne $G_{k,t}$, à savoir (j – i), dans le bruit global $\Theta + (j - i)$. La masse totale des pondérations n'est toutefois pas encore apparente dans la sommation.

[4] Les $r_{k'}^{(i,j)}$ ne doivent pas être confondus avec les taux de croissance monopériodiques de l'indice de ventes répétées : r_i ou r_j. La première quantité se calcule immédiatement avec les données, tandis que r_i et r_j sont les solutions du problème de minimisation, obtenues au terme du processus de calcul de l'indice.

Introduisons la fonction continue G(x) = x / (x + Θ). Sur l'intervalle temporel [0 ; +∞ [, G est strictement croissante de 0 à 1. Au fur et à mesure du passage du temps, la composante temporellement variable, mesurée par x = j – i, devient la principale source de bruit et G tend donc vers 1.
Les propriétés de cette fonction permettent de définir le concept de G-moyenne[5] des périodes de détention, pour toutes les ventes répétées pertinentes pour l'intervalle [t,t+1]. On la notera ζ^t et elle se définira par la relation[6] :

$$\sum_{i \le t < j} \sum_{k'} (j - i) / (\Theta + (j - i)) = \sum_{i \le t < j} \sum_{k'} G(j - i) = n^t G(\zeta^t)$$

La quantité $G(\zeta^t) = \zeta^t / (\zeta^t + \Theta)$ s'interprète, quant à elle, comme la moyenne arithmétique des proportions (j – i) / (Θ + (j – i)). Il s'agit donc d'une mesure de la contribution moyenne du bruit $G_{k,t}$ au bruit total, pour les transactions pertinentes pour [t,t+1].

La deuxième somme peut maintenant s'écrire :

$$\sum_{i \le t < j} \sum_{k'} (j - i) / (\Theta + (j - i)) \, r_{k'}^{(i,j)} = \sum_{i \le t < j} \sum_{k'} G(j - i) \, r_{k'}^{(i,j)}$$

et si l'on choisit comme nouvelle définition de ρ_t : $(n^t G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} G(j-i) r_{k'}^{(i,j)}$

elle devient simplement : $\sum_{i \le t < j} \sum_{k'} (j - i) / (\Theta + (j - i)) \, r_{k'}^{(i,j)} = n^t G(\zeta^t) \, \rho_t$

La masse totale des pondérations, qui manquait précédemment, est donc maintenant connue : elle vaut $n^t G(\zeta^t)$. ρ_t se définira donc toujours comme la moyenne

[5] On pourra se reporter à l'annexe 13 pour ce qui concerne le concept général de moyenne.
[6] le n^t représente le nombre d'éléments de la sommation : $\sum_{i \le t < j} \sum_{k'} 1 = n^t$

arithmétique des taux moyens mais, cette fois, pondérée par les coefficients G(j – i), dont la somme vaut $n^t G(\zeta^t)$.

On peut retrouver la formule simplifiée du paragraphe 3, en remarquant que dans ce cadre nous avons : $\Theta = 0$ $r_{k'}^{(i,j)} = r_{i,j}$ $G(x) = 1$

D'où : $\rho_t = (1 / n^t) \sum_{i \leq t < j} n_{i,j} r_{i,j}$

Annexe 8 : Reformulation de ρ_t dans le cas général et autres notations

Moyenne des taux moyens

Quand $\Theta = 0$ et $r_{k'}^{(i,j)} = r_{i,j}$ (cas simplifié du paragraphe 3), il a été démontré que ρ_t pouvait se réécrire :

$$\rho_t = (1/ \tau^t) * (\ln H_f(t) - \ln H_p(t)) = (I^t / n^t) * (\ln H_f(t) - \ln H_p(t))$$

τ^t, $H_f(t)$ et $H_p(t)$ correspondaient respectivement à la moyenne harmonique des périodes de détention, et aux moyennes géométriques des prix de revente (futur) et des prix d'achat (passé), calculées en supposant que toutes les transactions[7] étaient réalisées au niveau de l'indice H. Que deviennent cette relation et ces grandeurs dans le nouveau contexte ?

On a ici :
$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} \sum_{k'} G(j - i) \, r_{k'}^{(i,j)}$$
$$= (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} (\Theta + (j - i))^{-1} \sum_{k'} \ln(p_{k',j} / p_{k',i}).$$
$$= (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} (\Theta + (j - i))^{-1} \ln(\Pi_{k'} p_{k',j} / \Pi_{k'} p_{k',i})$$

Introduisons les moyennes géométriques des prix d'achat et des prix de revente[8] pour les couples de la classe (i,j) : $h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}}$ $\quad h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}}$

On obtient :
$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} L_{i,j} \ln (h_f^{(i,j)} / h_p^{(i,j)})$$
$$= (n^t G(\zeta^t))^{-1} \ln\Big[\Big(\prod_{i \le t < j} (h_f^{(i,j)})^{L_{i,j}} \Big) / \Big(\prod_{i \le t < j} (h_p^{(i,j)})^{L_{i,j}} \Big)\Big]$$

[7] Les transactions intervenant dans les calculs de τ^t, $H_f(t)$ et $H_p(t)$ sont les ventes répétées pertinentes pour [t,t+1]

[8] Un achat est un événement passé (p) et une revente un événement futur (f), par rapport au présent [t,t+1].

Puis, en introduisant les moyennes géométriques des $h_p^{(i,j)}$ et des $h_f^{(i,j)}$ pondérés par les $L_{i,j}$, dont le poids total vaut $I^t = \sum_{i \le t < j} L_{i,j}$:

$$H_p(t) = \left(\Pi_{i \le t < j} \left(h_p^{(i,j)} \right)^{L_{i,j}} \right)^{1 / I^t} \qquad H_f(t) = \left(\Pi_{i \le t < j} \left(h_f^{(i,j)} \right)^{L_{i,j}} \right)^{1 / I^t}$$

on peut alors écrire : $\rho_t = (I^t / (n^t G(\zeta^t))) * \ln [H_f(t) / H_p(t)]$

La formule du paragraphe 3 est donc globalement préservée. Deux modifications doivent être cependant pointées : un coefficient $G(\zeta^t)$ apparaît et les définitions de $H_p(t)$ et $H_f(t)$ sont un peu plus complexes.

<u>Prix moyens d'achat et de revente $H_p(t)$ et $H_f(t)$</u>

Ces deux expressions peuvent se réécrire :

$$\left[H_p(t) \right]^{I^t} = \Pi_{i \le t < j} \left(h_p^{(i,j)} \right)^{L_{i,j}} = \Pi_{i \le t < j} \left(\left(\Pi_{k'} p_{k',i} \right)^{1/n_{i,j}} \right)^{L_{i,j}}$$

$$\left[H_p(t) \right]^{I^t} = \Pi_{i \le t < j} \left(\Pi_{k'} p_{k',i} \right)^{1 / (\Theta + (j - i))}$$

$$\left[H_f(t) \right]^{I^t} = \Pi_{i \le t < j} \left(\Pi_{k'} p_{k',j} \right)^{1 / (\Theta + (j - i))}$$

Il s'agit donc toujours des moyennes géométriques des prix d'achat et de revente pour tous les couples[9] dont la période de détention englobe [t,t+1], mais cette fois les pondérations valent $1/(\Theta + (j - i))$. On peut également continuer à interpréter le terme $\ln[H_f(t) / H_p(t)]$ comme un rendement moyen.

[9] Cet échantillon est noté Spl^t.

Périodes et fréquences de détention

Les concepts de période moyenne de détention et de fréquence moyenne sont aussi généralisables.

On a : $I^t = \sum_{i \leq t < j} L_{i,j} = \sum_{i \leq t < j} \sum_{k'} (\Theta+(j-i))^{-1} = \sum_{i \leq t < j} \sum_{k'} G(j - i) * (1 / (j - i))$

et : $\sum_{i \leq t < j} \sum_{k'} (j – i) / (\Theta + (j – i)) = \sum_{i \leq t < j} \sum_{k'} G(j – i) = n^t G(\zeta^t)$

Le quotient $F^t = I^t / (n^t G(\zeta^t))$ est donc la moyenne arithmétique des fréquences de détention $1 / (j – i)$, pour les couples de Spl^t, pondérées par les $G(j – i)$.

Ou d'une manière équivalente, en posant $\tau^t = (F^t)^{-1} = (n^t G(\zeta^t)) / I^t$ on a :

$$(n^t G(\zeta^t)) / \tau^t = \sum_{i \leq t < j} \sum_{k'} G(j – i) * (1 / (j – i))$$

et τ^t s'interprète alors comme la moyenne harmonique des périodes de détention (j – i), dans Spl^t, pondérées par les G (j – i).

La formule de la moyenne des taux moyens se réécrit alors : $\rho_t = (1/\tau^t) \ln[H_f(t)/H_p(t)]$

Enfin, on peut définir grâce à $N = \sum_{i<j} n_{ij}$, $I = \sum_{i < j} L_{i,j}$ et la G-moyenne[10] ζ des périodes de détention, les équivalents de ces grandeurs pour l'échantillon pris dans son ensemble, à savoir :

$$F = I / (N G(\zeta)) \qquad \tau = F^{-1} = N G(\zeta) / I$$

[10] ζ est définie par la relation $\sum_{i < j} \sum_{k'} G(j – i) = N G(\zeta)$

Annexe 9 : Le lien entre ζ^t et τ^t

Deux notions de moyenne ont été introduites pour les temps de détention, la G-moyenne ζ^t et la moyenne harmonique τ^t. Il est alors légitime de se demander s'il existe un lien entre ces deux grandeurs. Elles sont définies par les relations suivantes:

- $n^t\, G(\zeta^t) = \sum_{i \le t < j} \sum_{k,} G(j - i)$
- $(n^t\, G(\zeta^t)) / \tau^t = \sum_{i \le t < j} \sum_{k,} G(j - i) * (1 / (j - i))$

La deuxième relation peut se réécrire : $\tau^t = (n^t\, G(\zeta^t)) / \sum_{i \le t < j} \sum_{k,} G(j-i)*(1/(j-i))$

La fonction G étant définie par $G(x) = x / (x + \Theta)$ on a alors :

$$G(\tau^t) = G(\zeta^t) \;/\; \left[G(\zeta^t) + (\Theta / n^t) \sum_{i \le t < j} \sum_{k,} G(j - i) * (1 / (j - i)) \right]$$

En utilisant la définition de ζ^t on obtient :

$$G(\tau^t) = G(\zeta^t) / \left[(1/n^t) \sum_{i \le t < j} \sum_{k,} G(j-i) + (\Theta/n^t) \sum_{i \le t < j} \sum_{k,} G(j-i)*(1/(j-i)) \right]$$

$$G(\tau^t) = G(\zeta^t) / \left[(1 / n^t) \sum_{i \le t < j} \sum_{k,} G(j - i) * (1 + \Theta / (j - i)) \right]$$

Or : $\quad 1 + \Theta / (j - i) = (\Theta + (j - i)) / (j - i) = [\, G(j - i) \,]^{-1}$

D'où : $\quad G(\tau^t) = G(\zeta^t) / \left[(1 / n^t) \sum_{i \le t < j} \sum_{k,} 1 \right] = G(\zeta^t)$

Comme la fonction G est strictement croissante, donc injective, on peut en déduire que $\tau^t = \zeta^t$. Les deux concepts de moyenne coïncident donc strictement. Cependant, ils ne sont pas pour autant redondants, car ils seront en général utilisés dans des contextes différents. Ainsi, ζ^t apparaîtra plutôt sous la forme $G(\zeta^t)$, proportion moyenne du bruit en provenance des tendances locales dans le bruit total, tandis que τ^t s'interprètera plus simplement comme une durée moyenne de détention.

Annexe 10 : La solution du problème de minimisation

En appliquant la formule présentée ci-dessus pour ρ_t, aux dérivées partielles de Φ on obtient :

$$\partial\Phi(R)/\partial r_t = 2\sum_{i \le t < j} L_{i,j}\,(r_i + \ldots + r_{j-1}) - 2\sum_{i \le t < j} L_{i,j}\{(1/n_{i,j})\sum_{k'} \ln(p_{k',j}/p_{k',i})\}$$
$$= 2\sum_{i \le t < j} L_{i,j}\,(r_i + \ldots + r_{j-1}) - 2\, n^t\, G(\zeta^t)\,\rho_t$$

La première somme peut se simplifier en utilisant la technique de l'annexe 2, faisant ainsi apparaître naturellement les quantités $I^{[t', t+1]}$:

$$\partial\Phi(R) / \partial r_t = 2 \sum_{t' \le t} I^{[t', t+1]}\, r_{t'} + 2 \sum_{t' > t} I^{[t, t'+1]}\, r_{t'} - 2\, n^t\, G(\zeta^t)\,\rho_t$$

La résolution du système d'équations $\{\partial\Phi(R) / \partial r_t = 0\}_{t=0,\ldots,T-1}$ fournit alors la solution du problème de minimisation. Elle peut s'exprimer de deux manières différentes.

Introduisons la matrice de l'information pertinente $\hat{I}$:

$$\hat{I} = \begin{pmatrix} I^{[0,1]} & I^{[0,2]} & I^{[0,3]} & & I^{[0,T]} \\ I^{[0,2]} & I^{[1,2]} & I^{[1,3]} & & I^{[1,T]} \\ I^{[0,3]} & I^{[1,3]} & I^{[2,3]} & & I^{[2,T]} \\ \vdots & & & & \\ I^{[0,T]} & I^{[1,T]} & I^{[2,T]} & & I^{[T-1,T]} \end{pmatrix}$$

et la matrice diagonale, notée η, dont les valeurs sont $n^0\, G(\zeta^0)$, … , $n^{T-1}\, G(\zeta^{T-1})$. Le système[11] d'équations et sa solution s'écriront :

$$\hat{I}\, R = \eta\, P \quad \Leftrightarrow \quad R = (\hat{I}^{-1} \eta)\ P$$

L'autre possibilité consiste à diviser les lignes du système par leur I^t respectif et à utiliser la formule $\rho_t = (I^t/(n^t\, G(\zeta^t)))\ \ln[H_f(t) / H_p(t)]$. Si l'on introduit alors le vecteur colonne Y, dont les composantes sont $\ln[H_f(t) / H_p(t)]$, et la matrice de diffusion J définie par :

$$J = \begin{pmatrix} 1 & I^{[0,2]}/I^0 & I^{[0,3]}/I^0 & & I^{[0,T]}/I^0 \\ I^{[0,2]}/I^1 & 1 & I^{[1,3]}/I^1 & & I^{[1,T]}/I^1 \\ I^{[0,3]}/I^2 & I^{[1,3]}/I^2 & 1 & & I^{[2,T]}/I^2 \\ \vdots & & & & \\ I^{[0,T]}/I^{T-1} & I^{[1,T]}/I^{T-1} & I^{[2,T]}/I^{T-1} & & 1 \end{pmatrix}$$

le problème devient simplement :

$$J\, R = Y \quad \Leftrightarrow \quad R = J^{-1}\, Y$$

[11] $R = (r_0, r_1, \ldots, r_{T-1})$ et $P = (\rho_0, \rho_1, \ldots, \rho_{T-1})$

Annexe 11 : Le cas de la situation BMN

Dans le paragraphe 3 du chapitre 1, la modélisation a été développée en utilisant deux hypothèses simplificatrices :

- toutes les transactions s'effectuent au niveau exact de l'indice H
- les imperfections du marché sont inexistantes mais les tendances locales peuvent s'exprimer.

La première hypothèse a été levée en introduisant les moyennes $h_p^{(i,j)} = (\Pi_{k'} p_{k',i})^{1/n_{i,j}}$ et $h_f^{(i,j)} = (\Pi_{k'} p_{k',j})^{1/n_{i,j}}$. Tout se passe alors comme si les achats de la classe (i,j) s'effectuaient tous au niveau $h_p^{(i,j)}$ et les reventes au niveau $h_f^{(i,j)}$.

Pour la deuxième, considérer que σ_N n'est plus nul a conduit à introduire le paramètre $\Theta = 2\sigma_N^2 / \sigma_G^2$, dont la fonction consiste à mesurer l'importance relative des deux sources de bruits (imperfections, tendances locales). Il intervient dans les différentes grandeurs, essentiellement sous la forme de pondérations. On passe par exemple pour τ^t d'une formule équipondérée dans le cas simplifié, à une formule plus complexe dans le cas général :

Cas simplifié : $$n^t / \tau^t = I^t = \sum_{i \le t < j} n_{i,j} / (j - i)$$

Cas général : $$(n^t G(\zeta^t)) / \tau^t = \sum_{i \le t < j} (n_{i,j} G(j - i)) / (j - i)$$

Certaines classes de ventes répétées seront ainsi plus importantes que d'autres dans les calculs, la contribution à la moyenne étant déterminée grâce à la fonction G et donc par Θ.

Le paramètre Θ varie entre 0 et $+\infty$. Le cas polaire $\Theta = 0$ correspond à la situation simplifiée[12]. Lorsque ce paramètre est dans l'intervalle] 0, $+\infty$ [, on se trouve dans

[12] Si l'on considère que la première hypothèse simplificatrice ne s'applique pas.

la situation générale de type Case, Shiller. Mais jusqu'ici, l'autre pole $\Theta = +\infty$, n'a pas été étudié. A quoi correspond-t-il ?

$\Theta = 2\sigma_N^2 / \sigma_G^2 = +\infty$ signifie en fait que σ_G^2 est négligeable devant $2\sigma_N^2$. En d'autres termes, on peut considérer que $\sigma_G = 0$ et que l'on se trouve dans une situation où les tendances locales (capturées par la marche aléatoire) sont absentes. Seules subsistent alors les imperfections du marché, décrites par le bruit blanc. La variance du résidu est alors homoscédastique et il s'agit en fait d'un contexte de type BMN[13]. Il n'est pas possible d'appliquer directement toutes les formules établies jusqu'à maintenant, car $+\infty$ n'est pas un nombre. Les adaptations nécessaires, si l'on souhaite conserver le même formalisme, sont les suivantes :

- Définition des $L_{i,j}$

  Pour chaque transaction la variance du résidu est identique ($= 2\sigma_N^2$) et elle ne dépend pas de la durée de détention, il en sera donc de même pour la quantité d'information d'une vente répétée. La mesure d'information étant définie de manière relative[14] et non pas absolue, il est alors souhaitable de choisir la plus simple des conventions. On normalisera donc la mesure du bruit en substituant à $2\sigma_N^2$ la valeur 1.

  Par conséquent la grandeur $L_{i,j}$ sera définie simplement par : $L_{i,j} = n_{i,j} / 1 = n_{i,j}$

- Définition de la fonction G

  G ne peut plus s'interpréter comme la proportion de bruit généré par les tendances locales dans le bruit total car cette source n'existe plus. Sa définition antérieure était $G(j - i) = (j - i) / (\Theta + (j - i))$, durée de détention divisée par la mesure de bruit. Le cadre BMN impose donc de choisir $G(j - i) = (j - i) / 1 = j - i$; la fonction G est donc simplement la fonction identité[15].

[13] BMN pour Bailey, Muth, Nourse (1963)

[14] C'est-à-dire à une constante multiplicative près.

[15] Dans le programme d'optimisation la quantité $(j - i) / (\Theta + (j - i))$ apparaît d'abord sous la forme $(j - i) / (2\sigma_N^2 + \sigma_G^2(j - i))$. Si $\sigma_G = 0$, elle devient $(j - i) / 2\sigma_N^2$ et par normalisation $(j - i)$. Le

- Durée de détention

  La variable ζ^t, G-moyenne des durées de détention, se définissait dans le contexte Case-Shiller par la relation : $\sum_{i \le t < j} \sum_{k'} G(j - i) = n^t \, G(\zeta^t)$

  Celle-ci devient dans le contexte BMN : $\sum_{i \le t < j} \sum_{k'} (j - i) = n^t \, \zeta^t$

  ζ^t n'est plus ici qu'une simple moyenne arithmétique équipondérée des durées de détention pour les ventes répétées pertinentes pour [t,t+1].

- Moyenne des taux moyens

  Ce taux se définit dans un premier temps par la formule :

$$\rho_t = (n^t \, \zeta^t)^{-1} \sum_{i \le t < j} \sum_{k'} (j - i) \, r_{k'}^{(i,j)}$$

  Les grandeurs $h_p^{(i,j)}$, $h_f^{(i,j)}$, $H_p(t)$ et $H_f(t)$ ne sont pas modifiées, tant dans leur définition que dans leur interprétation. On peut alors réécrire, comme dans le contexte Case-Shiller, le taux moyen sous la forme :

$$\rho_t = (I^t / (n^t \, \zeta^t)) * \ln [H_f(t) / H_p(t)]$$

  La relation $n_{i,j} = L_{i,j}$ implique que $I^t = n^t$; de là découle la formule simplifiée suivante :

$$\rho_t = (1 / \zeta^t) * \ln [H_f(t) / H_p(t)]$$

  On retrouve alors de manière évidente la relation $\zeta^t = \tau^t$

- Formule matricielle

  La relation $\hat{I}R = \eta \, P$ est conservée.

---

passage à la limite ($\Theta \to +\infty$), qui aboutit à $G(x) = 0$, n'est donc pas légitime. La définition correcte de la fonction G dans la situation BMN consiste donc bien à choisir $G(x) = x$.

Annexe 12 : Démonstration des formules générales pour l'échantillon benchmark

Limite de la suite (u_n)

$$u_n = \alpha /(\Theta + 1) + \alpha^2 /(\Theta + 2) + \alpha^3 /(\Theta + 3) + \ldots + \alpha^n /(\Theta + n)$$

$$= \left[\alpha^{\Theta + 1} /(\Theta + 1) + \alpha^{\Theta + 2} /(\Theta + 2) + \ldots + \alpha^{\Theta + n} /(\Theta + n) \right] / \alpha^{\Theta}$$

$$= \left[\sum_{i = 1, \ldots, \Theta + n} \alpha^i / i - (\alpha + \alpha^2 / 2 + \ldots + \alpha^{\Theta} / \Theta) \right] / \alpha^{\Theta}$$

D'où : $$\ell = \text{Lim } u_n = - \left[\ln (1 - \alpha) + \alpha + \alpha^2 / 2 + \ldots + \alpha^{\Theta} / \Theta \right] / \alpha^{\Theta}$$

Des dénominateurs non entiers auraient empêché l'usage de cette technique, choisir des valeurs entières pour Θ est donc une simplification utile.

Quantité d'information I

$$I = K' \sum_{k = 1,\ldots,T} [(T - k + 1) / (\Theta + k)] \alpha^k \qquad \text{(somme sur les diagonales)}$$

$$= K' (T + \Theta + 1) \sum_{k = 1,\ldots,T} \alpha^k / (\Theta + k) - K' \sum_{k = 1,\ldots,T} \alpha^k$$

$$= K' (T + \Theta + 1) u_T - K' (\alpha /(1 - \alpha)) * (1 - \alpha^T)$$

$$= K' [(T + \Theta + 1) u_T - T \pi]$$

Proportion d'information manquante

$$\mu = (K'T \ell - I) / K'T \ell$$

$$= [T \ell - (T + \Theta + 1) u_T + T \pi] / T \ell$$

$$= [\ell + \pi - (T + \Theta + 1) (u_T / T)] / \ell$$

Annexe 13 : La généralisation du concept de moyenne

La méthode la plus couramment employée pour calculer une moyenne consiste à appliquer une formule arithmétique et, dans une moindre mesure, une formule géométrique ou harmonique. Si à première vue ces calculs peuvent sembler différents, il n'en est rien car ils relèvent tous d'une seule et même structure.

Définition

Soit F une fonction continue de I vers J (deux intervalles réels), strictement croissante (ou décroissante). La moyenne des nombres $\{x_1 , x_2 , \dots , x_n\}$ pondérés par les poids $(\alpha_1 , \alpha_2 , \dots , \alpha_n)$ est le nombre réel vérifiant :

$$\alpha\, F(X) = \alpha_1 F(x_1) + \alpha_2 F(x_2) + \dots + \alpha_n F(x_n) \text{ avec } \alpha = \sum_{i=1,\dots,n} \alpha_i$$

Pour chaque fonction F on obtient ainsi une moyenne particulière.

Exemples

Moyenne arithmétique : $F(x) = x$

$$\alpha\, X = \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_n x_n \quad \Leftrightarrow \quad X = (\alpha_1 x_1 + \dots + \alpha_n x_n) / \alpha$$

Moyenne géométrique : $F(x) = \ln(x)$

$$\alpha \ln(X) = \alpha_1 \ln(x_1) + \dots + \alpha_n \ln(x_n) \Leftrightarrow \quad X = (x_1^{\alpha_1} \dots x_n^{\alpha_n})^{(1/\alpha)}$$

Moyenne harmonique : $F(x) = 1/x$

$$\alpha / X = \alpha_1 / x_1 + \dots + \alpha_n / x_n \quad \Leftrightarrow \quad X = \alpha * (\alpha_1 / x_1 + \dots + \alpha_n / x_n)^{-1}$$

G-moyenne

La fonction continue G(x) = x / (x + Θ) est strictement croissante sur l'intervalle temporel [0;+∞ [et varie de 0 à 1. Elle permet de définir dans le modèle des ventes répétées la grandeur ζ^t :

- ζ^t est la G-moyenne des périodes de détention pour les ventes répétées pertinentes pour l'intervalle [t,t+1]. Elle se définit[16] par la relation $\sum_{i \leq t < j} \sum_{k'} G(j-i) = n^t G(\zeta^t)$

On a alors : $G(\zeta^t) = (1/ n^t) \sum_{i \leq t < j} \sum_{k'} (j - i) / (\Theta + (j - i)) = \zeta^t / (\zeta^t + \Theta)$

$G(\zeta^t)$ est donc la moyenne arithmétique équipondérée des proportions (j-i) / (Θ+(j-i)). Elle mesure la contribution moyenne du bruit $G_{k,t}$ au bruit total, pour les données pertinentes pour l'intervalle [t,t+1] .

[16] On rappelle ici que pour un couple (i,j) donné et un temps t fixé, avec $i \leq t < j$, le compteur k' parcourt l'ensemble des $n_{i,j}$ ventes répétées de la cellule (i,j).

Annexe 14 : Reformulation de $H_p(t)$ et $H_f(t)$ en fonction de l'indice de prix H

L'expression de $H_p(t)$ est :

$$\left[H_p(t)\right]^{I^t} = \Pi_{i \le t < j} \left[\left(\Pi_{k'} p_{k',i} \right)^{1/(\Theta+(j-i))} \right] = \Pi_{i=0,\ldots,t} \left[\Pi_{j>t} \left(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})} \right) \right]$$

Pour chaque i entre 0 et t, les prix à l'intérieur du crochet forment un sous-ensemble de F_i , comme représenté dans la figure 3 (pour i = 2). Ce terme est quasiment une moyenne géométrique des prix, pondérés par les contributions informationnelles. La masse totale, qui n'apparaît pas encore directement, est :

$$\sum_{j>t} \left(\sum_{k'} \inf(p_{k',i}) \right) = \sum_{j>t} \sum_{k'} (\Theta + (j-i))^{-1} = \sum_{j>t} L_{i,j}$$

$$= L_{i,t+1} + L_{i,t+2} + \ldots + L_{i,T} = B^t_i$$

Figure 3: Sous-ensemble de F_2 utilisé dans le calcul de $H_p(t)$

	0	1	2	…	t	t + 1	…	T – 1	T
0									
1									
2									
⋮									
t									
t + 1									
⋮									
T – 1									
T									

Il n'est pas évident a priori que les deux moyennes $\left[\Pi_{j>t} \left(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})} \right) \right]^{1/B^t_i}$ et h_t soient égales, car la moyenne dans un sous-ensemble peut varier légèrement autour de la moyenne globale[17].

[17] Et cela d'autant plus que le sous-échantillon est petit

Cependant on peut raisonnablement écrire :

$$[\Pi_{j>t}(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})})]^{1/B^t_i} = h_i \exp(\upsilon(i,t)) \quad \text{avec } \upsilon(i,t) \approx 0 \text{ et } E[\upsilon(i,t)] = 0.$$

La fonction de $\upsilon(i,t)$ est de capturer la variabilité de la moyenne des prix lorsque celle-ci est calculée sur un sous-ensemble de F_i.

En utilisant ces notations pour tous les i entre 0 et t on a :

$$[H_p(t)]^{I^t} = \Pi_{i=0,\ldots,t}[\Pi_{j>t}(\Pi_{k'} p_{k',i}^{\inf(p_{k',i})})] = \Pi_{i=0,\ldots,t}[h_i \exp(\upsilon(i,t))]^{B^t_i}$$

$$= [\Pi_{i=0,\ldots,t} h_i^{B^t_i}] \exp(B^t_0 \upsilon(0,t) + B^t_1 \upsilon(1,t) + \ldots + B^t_t \upsilon(t,t))$$

$$[H_p(t)] = [\Pi_{i=0,\ldots,t} h_i^{B^t_i}]^{1/I^t} \exp((B^t_0/I^t)\upsilon(0,t) + \ldots + (B^t_t/I^t)\upsilon(t,t))$$

et en notant $\upsilon^t = (B^t_0 / I^t)\upsilon(0,t) + \ldots + (B^t_t / I^t)\upsilon(t,t)$ on obtient finalement :

$$[H_p(t)] = [\Pi_{i=0,\ldots,t} h_i^{B^t_i}]^{1/I^t} \exp(\upsilon^t)$$

De manière similaire pour le futur, c'est à dire côté vente, on peut facilement établir que :

$$[H_f(t)] = [\Pi_{j=t+1,\ldots,T} h_j^{S^t_j}]^{1/I^t} \exp(\nu^t)$$

Les quantités ν^t et υ^t sont telles que : $\nu^t \approx 0$ $\upsilon^t \approx 0$ $E[\nu^t] = 0$ $E[\upsilon^t] = 0$.

Annexe 15 : Démonstration des propositions du chapitre 3, paragraphe 2.1

Deuxième formule de la proposition 3

$$\begin{aligned} Tr(\hat{I}) - Tr_{+1}(\hat{I}) &= \sum_{t=0,\ldots,T-1} I^{t} - \sum_{t=0,\ldots,T-2} I^{[t,\,t+2]} \\ &= \sum_{t=0,\ldots,T-2} (I^{t} - I^{[0,\,t+2]}) + I^{T-1} \\ &= \sum_{t=0,\ldots,T-2} \sum_{i=0,\ldots,t} L_{i,\,t+1} + \sum_{i=0,\ldots,T-1} L_{i,\,T} \\ &= \sum_{t=0,\ldots,T-1} \sum_{i=0,\ldots,t} L_{i,\,t+1} = I \end{aligned}$$

Troisième formule de la proposition 3

Pour n^t la relation est moins évidente que pour I^t mais on peut toutefois la mettre en évidence en réutilisant la formule établie dans l'annexe 2 :

$$\sum_{i \le t < j} L_{i,j}\,(r_i + \ldots + r_{j-1}) = \sum_{t' \le t} I^{[t',\,t+1]}\, r_{t'} + \sum_{t' > t} I^{[t,\,t'+1]}\, r_{t'}$$

En choisissant $r_0 = r_1 = \ldots = r_{T-1} = 1$ on a :

$$\begin{aligned} \sum_{i \le t < j} (j - i)\, L_{i,j} &= \sum_{t' \le t} I^{[t',\,t+1]} + \sum_{t' > t} I^{[t,\,t'+1]} \\ \sum_{i \le t < j} \sum_{k'} G(j - i) &= \sum_{t' \le t} I^{[t',\,t+1]} + \sum_{t' > t} I^{[t,\,t'+1]} \end{aligned}$$

D'où : $$\sum_{t' \le t} I^{[t',\,t+1]} + \sum_{t' > t} I^{[t,\,t'+1]} = n^t\, G(\zeta^t)$$

Annexe 16 : Variance et covariance pour le vecteur P

Pour établir ces résultats on utilise la dynamique de $\rho_{i,j}$: $\mathcal{N}(r^{(i,j)} ; \sigma_G^2 / (L_{i,j}(j-i)^2))$. On rappelle de plus que les variables aléatoires $\rho_{i,j}$ sont indépendantes entre elles.

Variance

D'une manière générale on a pour X et Y indépendantes : $V(X + Y) = V(X) + V(Y)$ et si a est une constante : $V(aX) = a^2 V(X)$

De $\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j - i) n_{i,j} \rho_{i,j}$ on déduit alors :

$$
\begin{aligned}
V(\rho_t) &= (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} G^2(j - i) (n_{i,j})^2 \sigma_G^2 / (L_{i,j} (j - i)^2) \\
&= (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} (n_{i,j})^2 \sigma_G^2 / (L_{i,j} (\Theta + j - i)^2) \\
&= (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} (n_{i,j})^2 \sigma_G^2 / (n_{i,j} (\Theta + j - i)) \\
&= (n^t G(\zeta^t))^{-2} \sum_{i \le t < j} n_{i,j} \sigma_G^2 / (\Theta + j - i) \\
&= (\sigma_G / (n^t G(\zeta^t)))^2 \sum_{i \le t < j} L_{i,j} \\
&= \sigma^2_G / ((\tau^t)^2 I^t)
\end{aligned}
$$

Covariance

Pour t, t' nous avons :

$$\rho_t = (n^t G(\zeta^t))^{-1} \sum_{i \le t < j} G(j-i) n_{i,j} \rho_{i,j} \quad \text{et} \quad \rho_{t'} = (n^{t'} G(\zeta^{t'}))^{-1} \sum_{i \le t' < j} G(j-i) n_{i,j} \rho_{i,j}$$

Si les dates t et t' diffèrent, les grandeurs ρ_t et $\rho_{t'}$ ne sont pas pour autant indépendantes. Certaines classes (i,j) sont, en effet, dans les deux sous-échantillons d'effectifs pertinents sur lesquels reposent les calculs de ρ_t et $\rho_{t'}$.

Les moyennes $\rho_{i,j}$ des rentabilités dans les classes (i,j) sont, par contre, indépendantes entre elles. Si l'on prend deux classes (i,j) et (i',j'), la covariance des deux moyennes $\rho_{i,j}$ et $\rho_{i',j'}$ est donc nulle. Sauf s'il s'agit des mêmes classes (i = i' et j = j'). Dans ce cas la covariance n'est rien d'autre que la variance de la rentabilité de cette classe, c'est-à-dire $\sigma_G^2 / (L_{i,j} (j - i)^2)$. En utilisant les propriétés de bilinéarité de la covariance, on peut alors en déduire que :

$$\begin{aligned}
\mathrm{Cov}(\rho_t ; \rho_{t'}) &= (n^{t'} n^t G(\zeta^{t'}) G(\zeta^t))^{-1} \sum_{i \le t' < t < j} (G(j-i)\, n_{i,j})^2 \sigma_G^2 / (L_{i,j} (j-i)^2) \\
&= (n^{t'} n^t G(\zeta^{t'}) G(\zeta^t))^{-1} \sigma_G^2 \sum_{i \le t' < t < j} L_{i,j} \\
&= (n^{t'} n^t G(\zeta^{t'}) G(\zeta^t))^{-1} \sigma_G^2 I^{[t', t+1]} \\
&= \sigma_G^2 I^{[t', t+1]} / (\tau^t \tau^{t'} I^t I^{t'})
\end{aligned}$$

Pour t = t' on retrouve bien sur la formule de la variance : $\sigma^2_G / ((\tau^t)^2 I^t)$

Annexe 17 : La matrice de variance-covariance de LInd

Le produit matriciel A'^{-1} Î A^{-1} peut devenir plus lisible en utilisant quelques éléments d'algèbre linéaire. Supposons, à titre d'exemple, que la matrice A soit de dimension 4. On a dans ce cas :

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{pmatrix} \qquad A^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}$$

$$A'^{-1} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Les matrices A'^{-1} et A^{-1} sont des matrices ne comportant que des +1 sur leur diagonale principale et des -1 sur la diagonale supérieure pour A'^{-1}, la diagonale inférieure pour A^{-1}. Les autres coordonnées sont nulles. Ce résultat est général et il ne dépend pas de la dimension de la matrice.

Appliquer à gauche la matrice A'^{-1} à Î revient à soustraire à chaque ligne "i" de Î, la ligne suivante "i + 1" ; à l'exception de la dernière ligne qui reste inchangée. On applique ensuite au résultat obtenu la matrice A^{-1} à droite. Cette opération revient à retrancher à chaque colonne "j" la colonne "j + 1" ; la dernière restant inchangée.

Indexons la matrice $\hat{I}$ par des coordonnées (p,q), où $1 \leq p, q \leq T$. Ces opérations élémentaires sur les lignes et les colonnes de $\hat{I}$ produisent quatre types de résultats pour les valeurs de la matrice $A'^{-1}\, \hat{I}\, A^{-1}$ (on notera $b_{p,q}$ ses éléments).

- Pour $1 \leq p, q < T$: $\hat{I}_{p,q} - (\hat{I}_{p+1,q} + \hat{I}_{p,q+1}) + \hat{I}_{p+1,q+1}$
- Pour $1 \leq q < T$: $\hat{I}_{T,q} - \hat{I}_{T,q+1}$
- Pour $1 \leq p < T$: $\hat{I}_{p,T} - \hat{I}_{p+1,T}$
- Pour $p = q = T$: $\hat{I}_{T,T}$

Ces résultats sont très proches de ceux énoncés dans la proposition 2 du paragraphe 2.1 (chapitre 3). Ils doivent donc pouvoir se simplifier.

Cas 1 : $1 \leq p < q < T$

En décalant les indices, la proposition 2 affirme que pour ces valeurs de p et q on a :

$$\hat{I}_{p,q} = \hat{I}_{p+1,q} + \hat{I}_{p,q+1} - \hat{I}_{p+1,q+1} - L_{p,q}$$

D'où : $$b_{p,q} = - L_{p,q}$$

Cas 2 : $1 \leq q < p < T$

En utilisant la propriété de symétrie des coefficients de $\hat{I}$ l'expression se réécrit

$$b_{p,q} = \hat{I}_{q,p} - (\hat{I}_{q+1,p} + \hat{I}_{q,p+1}) + \hat{I}_{q+1,p+1}$$

On se retrouve dans le Cas 1 mais avec des coordonnées permutées, d'où :

$$b_{p,q} = - L_{q,p}$$

Cas 3 : $1 \leq q = p < T$

La proposition 2 ne s'applique plus, il faut raisonner directement :

$$b_{p,p} = \hat{I}_{p,p} - (\hat{I}_{p+1,p} + \hat{I}_{p,p+1}) + \hat{I}_{p+1,p+1}$$

$$b_{p,p} = \hat{I}_{p,p} - 2\, \hat{I}_{p,p+1} + \hat{I}_{p+1,p+1}$$

$$b_{p,p} = (L_{p,p+1} + L_{p,p+2} + \ldots + L_{p,T}) + (L_{0,p} + L_{1,p} + \ldots + L_{p-1,p})$$

$$b_{p,p} = B^{p}_{p} + S^{p-1}_{p}$$

Cas 4 : p = T et $1 \leq q < T$

$$b_{T,q} = \hat{I}_{T,q} - \hat{I}_{T,q+1} = \hat{I}_{q,T} - \hat{I}_{q+1,T}$$

$$b_{T,q} = -L_{q,T}$$

Cas 5 : q = T et $1 \leq p < T$

$$b_{p,T} = \hat{I}_{p,T} - \hat{I}_{p+1,T}$$

$$b_{p,T} = -L_{p,T}$$

Cas 6 : $$b_{T,T} = \hat{I}_{T,T} = L_{0,T} + L_{1,T} + \ldots + L_{T-1,T} = S^{T-1}_{T}$$

De manière synthétique, les cœfficients $b_{p,q}$ de la matrice $A'^{-1}\ \hat{I}\ A^{-1}$ sont de deux types :

- Pour $p \neq q$, ils sont égaux à l'opposé de $L_{p,q}$ (ou $L_{q,p}$ si $p > q$) , c'est-à-dire à l'opposé de la quantité d'information fournie par les ventes répétées de la classe (t_p , t_q).

- Quand p = q et p < T (cas 3), la valeur obtenue, $B^{p}_{p} + S^{p-1}_{p}$, est d'une nature très différente. Il s'agit de la quantité d'information fournie par toutes les transactions de la date p, indépendamment de leur sens (achat ou revente). Cette interprétation vaut aussi pour p = q = T puisqu'à l'intérieur de l'échantillon d'estimation les seules transactions possibles à la date T sont des reventes (d'où le S^{T-1}_{T} dans cas 6).

Annexe 18 : Reformulation de $\mathcal{V}$(LInd)

La matrice $\mathcal{V}(\text{LInd})$ peut se réécrire :

$$\mathcal{V}(\text{LInd}) = \sigma_G^2 \left[\mathcal{T} - (\mathcal{L} + \mathcal{L}') \right]^{-1}$$

$$= \sigma_G^2 \left[(Id - (\mathcal{L} + \mathcal{L}') \mathcal{T}^{-1}) \mathcal{T} \right]^{-1}$$

$$= \sigma_G^2 \, \mathcal{T}^{-1} \left[Id - (\mathcal{L} + \mathcal{L}') \mathcal{T}^{-1} \right]^{-1}$$

Pour calculer explicitement l'inverse de la matrice entre crochets, nous allons avoir recours à la théorie des espaces vectoriels normés de matrices. Pour cela, on munit l'espace $\mathbb{R}^T$ des vecteurs de dimension T de la norme vectorielle suivante :

$$\| (x_1, x_2, \ldots, x_T) \|_1 = \sum_{i=1,\ldots,T} | x_i |$$

Cette norme de $\mathbb{R}^T$ induit sur l'espace des matrices carrées de dimension T une norme matricielle définie par :

$$\| M \|_1 = \| (m_{i,j})_{i,j=1,\ldots,T} \|_1 = \text{Max}_{j=1,\ldots,T} \sum_{i=1,\ldots,T} | m_{i,j} |$$

Cette « norme 1 » pour une matrice M correspond simplement à la valeur maximale que l'on obtient en sommant sur chaque colonne les coefficients $m_{i,j}$, en valeurs absolues.

La première étape du raisonnement consiste à prouver que la norme de la matrice $(\mathcal{L} + \mathcal{L}')\mathcal{T}^{-1}$ est strictement plus petite que 1. Pour cela, nous pouvons remarquer que multiplier à droite par l'inverse de la matrice diagonale $\mathcal{T}$ revient en fait à diviser les colonnes de $\mathcal{L} + \mathcal{L}'$ par les valeurs situées sur la diagonale de $\mathcal{T}$.

Plus précisément, prenons la p^{ème} colonne de $\mathcal{L} + \mathcal{L}'$. Elle s'écrit :

$$(L_{1,p} , L_{2,p} , \dots , L_{p-1,p} , 0, L_{p ,p+1} , \dots , L_{p,T})'$$

Le p^{ème} élément de la diagonale de $\mathcal{T}$ vaut :

$$b_{p,p} = (L_{p,p+1} + L_{p,p+2} + \dots + L_{p,T}) + (L_{0,p} + L_{1,p} + \dots + L_{p-1,p})$$

On retrouve tous les éléments de la p^{ème} colonne de $\mathcal{L} + \mathcal{L}'$ dans le terme utilisé comme dénominateur, la quantité $L_{0,p}$ en plus. Les $L_{i,j}$ étant positifs, la somme des valeurs absolues des éléments de la p^{ème} colonne de $(\mathcal{L} + \mathcal{L}')\mathcal{T}^{-1}$ est, par conséquent, inférieure strictement à 1 (sous réserve que $L_{0,p} \neq 0$). Ce raisonnement étant valable pour chaque colonne, y compris la dernière, on peut donc affirmer que $\| (\mathcal{L}+\mathcal{L}')\mathcal{T}^{-1} \|_1$ est bien strictement plus petit que 1 (à condition que la ligne des $L_{0,j}$ du tableau des $\{L_{i,j}\}$ ne comporte pas de zéros). L'intérêt de ce résultat apparaît avec la proposition suivante :

Proposition :
L'espace des matrices carrées de dimension T est muni d'une norme matricielle $\| \ \|$

Si la matrice M est telle que : $\| M \| < 1$

Alors la série de matrice $S = \sum_0^{+\infty} M^i$ est convergente

et $S (Id – M) = (Id – M) S = Id$

(l'inverse de la matrice Id – M est la matrice S)

Si l'on applique cette proposition au terme entre crochets dans l'expression de $\mathcal{V}$(LInd), on obtient alors :

$$\mathcal{V}(\text{LInd}) = \sigma_G^2 \, \mathcal{T}^{-1} \left[Id – (\mathcal{L} + \mathcal{L}')\mathcal{T}^{-1} \right]^{-1} = \sigma_G^2 \, \mathcal{T}^{-1} \sum_0^{+\infty} \left[(\mathcal{L} + \mathcal{L}')\mathcal{T}^{-1} \right]^i$$

Annexe 19 : Le lien entre les périodes de détention moyennes $\tau^t(T_1)$ et $\tau^t(T_2)$.

Par définition : $I^t(T_2\backslash T_1) = \sum_{i \le t < T_1 < j \le T_2} L_{i,j} = \sum_{i \le t < T_1 < j \le T_2} \sum_{k'} G(j-i) * (1 / (j-i))$

$I^t(T_2 \setminus T_1)$ est presque égal à la moyenne arithmétique des fréquences $1 / (j - i)$, pondérées par les $G(j - i)$. Il ne manque, encore une fois, que la masse totale des pondérations, c'est-à-dire :

$$\sum_{i \le t < T_1 < j \le T_2} \sum_{k'} G(j - i) = n^t(T_2 \setminus T_1) G(\zeta^t(T_2 \setminus T_1))$$

$\zeta^t(T_2\backslash T_1)$ est ici la G-moyenne des périodes de détention pour les nouvelles ventes répétées et la quantité $G(\zeta^t(T_2\backslash T_1))$ s'interprète toujours comme le niveau moyen du bruit $G_{k,t}$ dans le bruit total. Comme dans la situation basique $I^t(T_2\backslash T_1) / [n^t(T_2\backslash T_1) G(\zeta^t(T_2\backslash T_1))]$ correspond à une fréquence moyenne, notée $F^t(T_2\backslash T_1)$, et son inverse, noté $\tau^t(T_2\backslash T_1)$, à une moyenne harmonique des périodes de détention.

De : $I^t(T_2 \setminus T_1) = [n^t(T_2 \setminus T_1) G(\zeta^t(T_2 \setminus T_1))] / \tau^t(T_2 \setminus T_1)$

et : $I^t(T_2) = I^t(T_1) + I^t(T_2 \setminus T_1)$

On déduit alors que :

$$\left[n^t(T_2)\, G(\zeta^t(T_2))\right]/\tau^t(T_2) = \left[n^t(T_1) G(\zeta^t(T_1))\right]/\tau^t(T_1) + \left[n^t(T_2\backslash T_1) G(\zeta^t(T_2\backslash T_1))\right]/\tau^t(T_2\backslash T_1)$$

La relation $n^t(T_2)G(\zeta^t(T_2)) = n^t(T_1)G(\zeta^t(T_1)) + n^t(T_2\backslash T_1)G(\zeta^t(T_2\backslash T_1))$ permet alors d'affirmer que $\tau^t(T_2)$ est simplement la moyenne harmonique pondérée de $\tau^t(T_1)$ et $\tau^t(T_2\backslash T_1)$.

Annexe 20 : Réversibilité pour ρ_t.

Pour $t < T_1$ on a :

$$\rho_t(T_2) = \left[I^t(T_2) / \left(n^t(T_2)\, G(\zeta^t(T_2)) \right) \right] * \ln \left[H_f(t,T_2) / H_p(t,T_2) \right]$$

$$= \left[I^t(T_2) / \left(n^t(T_2)\, G(\zeta^t(T_2)) \right) \right] * \left[\ln H_f(t,T_2) - \ln H_p(t,T_2) \right]$$

$$= \left[I^t(T_1) \ln H_f(t,T_1) + I^t(T_2 \backslash T_1) \ln H_f(t, T_2 \backslash T_1) \right] / \left[I^t(T_2)\, \tau^t(T_2) \right]$$

$$- \left[I^t(T_1) \ln H_p(t,T_1) + I^t(T_2 \backslash T_1) \ln H_p(t, T_2 \backslash T_1) \right] / \left[I^t(T_2)\, \tau^t(T_2) \right]$$

$$= \left[(1 / \tau^t(T_1)) * \left(\ln H_f(t,T_1) - \ln H_p(t,T_1) \right) \right] * \left[\left(I^t(T_1)\, \tau^t(T_1) \right) / \left(I^t(T_2)\, \tau^t(T_2) \right) \right]$$

$$+ \left[(1/\tau^t(T_2 \backslash T_1)) * \left(\ln H_f(t,T_2 \backslash T_1) - \ln H_p(t,T_2 \backslash T_1) \right) \right] * \left[\left(I^t(T_2 \backslash T_1)\, \tau^t(T_2 \backslash T_1) \right) / \left(I^t(T_2)\, \tau^t(T_2) \right) \right]$$

On reconnaît dans le premier crochet l'expression de $\rho_t(T_1)$. De plus, on peut facilement prouver que l'expression dans le troisième, $(1/\tau^t(T_2 \backslash T_1)) * \left(\ln H_f(t,T_2 \backslash T_1) - \ln H_p(t,T_2 \backslash T_1) \right)$, est égale à $\left[n^t(T_2 \backslash T_1)\, G(\zeta^t(T_2 \backslash T_1)) \right]^{-1} \sum_{i \le t < T_1 < j \le T_2} \sum_{k'} G(j - i)\, r_{k'}^{(i,j)}$. Il s'agit donc là simplement de la moyenne pondérée des taux moyens $r_{k'}^{(i,j)}$, pour les nouvelles ventes répétées. On la notera bien sûr par $\rho_t(T_2 \backslash T_1)$.

La formule de réversibilité pour ρ_t s'écrit donc, pour $t < T_1$:

$$\rho_t(T_2) = \left[I^t(T_1) / I^t(T_2) \right] \left[\tau^t(T_1) / \tau^t(T_2) \right] \rho_t(T_1) + \left[I^t(T_2 \backslash T_1) / I^t(T_2) \right] \left[\tau^t(T_2 \backslash T_1) / \tau^t(T_2) \right] \rho_t(T_2 \backslash T_1)$$

Annexe 21 : Loi de réversibilité dans un cas simplifié

A partir d'un échantillon ω_0 de ventes répétées, on construit un indice sur $[0,T_1]$ où interviennent notamment les quantités $\hat{I}(T_1)$ et $R(T_1)$. Puis, en utilisant le benchmark exponentiel étalonné sur $[0,T_1]$, on simule la matrice $\hat{I}(T_2\backslash T_1)$. Les quantités $\hat{I}(T_1)$, $R(T_1)$, $\hat{I}(T_2\backslash T_1)$, et donc aussi $\hat{I}(T_2)$, sont fixes.

Vecteur $R(T_2)$

Dans la formule de réversibilité $R(T_2) = [\,\hat{I}(T_2)\,]^{-1}\,[\hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)R(T_2\backslash T_1)]$, la seule composante aléatoire est $R(T_2\backslash T_1)$ qui suit une loi $\mathcal{N}(\,R_{hyp}\,;\,\sigma_G^2\,\hat{I}(T_2\backslash T_1)^{-1}\,)$. Le vecteur $R(T_2)$, lié linéairement à $R(T_2\backslash T_1)$, est donc aussi un vecteur gaussien dont l'espérance et la variance peuvent se calculer de la manière suivante :

$E[R(T_2)] = [\hat{I}(T_2)]^{-1}\,[\hat{I}(T_1)E[R(T_1)] + \hat{I}(T_2\backslash T_1)E[R(T_2\backslash T_1)]]$ — $Y=MX \Rightarrow E[Y]=ME[X]$

$E[R(T_2)] = [\hat{I}(T_2)]^{-1}\,[\hat{I}(T_1)R(T_1) + \hat{I}(T_2\backslash T_1)E[R(T_2\backslash T_1)]]$ — $R(T_1)$ non aléatoire

$E[R(T_2)] = [\,\hat{I}(T_2)\,]^{-1}\,[\hat{I}(T_1)\,R(T_1) + \hat{I}(T_2\backslash T_1)\,R_{hyp}]$ — $E[R(T_2\backslash T_1)] = R_{hyp}$

$E[R(T_2)] = [\hat{I}(T_2)]^{-1}\,[(\hat{I}(T_1)+\hat{I}(T_2\backslash T_1))R(T_1) + \hat{I}(T_2\backslash T_1)R_{hyp}(T_1;T_2)]$ — $R_{hyp} = R(T_1) + R_{hyp}(T_1;T_2)$

$E[R(T_2)] = R(T_1) + [\,[\,\hat{I}(T_2)]^{-1}\,\hat{I}(T_2\backslash T_1)]\,R_{hyp}(T_1;T_2)$ — $\hat{I}(T_2) = \hat{I}(T_1) + \hat{I}(T_2\backslash T_1)$

$\mathcal{V}[R(T_2)] = \mathcal{V}[\,[\hat{I}(T_2)\,]^{-1}\,\hat{I}(T_2\backslash T_1)\,R(T_2\backslash T_1)]$ — $R(T_1)$ non aléatoire

$\mathcal{V}[R(T_2)] = [\hat{I}(T_2)]^{-1}\,\hat{I}(T_2\backslash T_1)\mathcal{V}[R(T_2\backslash T_1)]\,(\hat{I}(T_2\backslash T_1))'([\hat{I}(T_2)]^{-1})'$ — $Y=MX \Rightarrow \mathcal{V}Y=M(\mathcal{V}X)M'$

$\mathcal{V}[R(T_2)] = [\hat{I}(T_2)]^{-1}\,\hat{I}(T_2\backslash T_1)\,\mathcal{V}[R(T_2\backslash T_1)]\;\hat{I}(T_2\backslash T_1)\,[\hat{I}(T_2)]^{-1}$ — matrices symétriques

$\mathcal{V}[R(T_2)] = \sigma_G^2\,[\hat{I}(T_2)]^{-1}\,\hat{I}(T_2\backslash T_1)\,\hat{I}(T_2\backslash T_1)^{-1}\,\hat{I}(T_2\backslash T_1)\,[\hat{I}(T_2)]^{-1}$ — $\mathcal{V}[R(T_2\backslash T_1)] = \sigma_G^2\hat{I}(T_2\backslash T_1)^{-1}$

$\mathcal{V}[R(T_2)] = \sigma_G^2\,[\,[\,\hat{I}(T_2)]^{-1}\,\hat{I}(T_2\backslash T_1)]\,[\,\hat{I}(T_2)]^{-1}$

Vecteur LInd(T_2)

Le vecteur LInd(T_2) est lié à R(T_2) par la formule LInd(T_2) = A(T_2) R(T_2). La matrice A(T_2) est de dimension T_2, elle est constituée de 1 sur sa diagonale et en dessous, de 0 ailleurs. La relation étant linéaire le vecteur LInd(T_2) est donc lui aussi un vecteur gaussien.

Dans le contexte du paragraphe 3.8.4, chapitre 3, nous avons supposé que $R_{hyp}(T_1;T_2)$ était un vecteur nul. On a dans ce cas E[R(T_2)] = R(T_1), d'où l'on déduit immédiatement que l'espérance du vecteur LInd(T_2), pour ses T_1 premières coordonnées, est simplement LInd(T_1).

On obtient la matrice de variance-covariance de LInd(T_2) en appliquant la formule classique :

$$\mathcal{V}[\text{LInd}(T_2)] = \sigma_G^2\ A(T_2)\ \left[[\ \hat{I}(T_2)]^{-1}\ \hat{I}(T_2 \backslash T_1)\right]\ [\ \hat{I}(T_2)]^{-1}\ [A(T_2)]'$$

Pour t = 1,…,T_1 $\text{LInd}_t(T_2) - \text{LInd}_t(T_1) = \text{Ln}(\text{Ind}_t(T_2) / \text{Ind}_t(T_1))$ suit donc une loi normale centrée, dont la variance v(t) est fournie par le t^{eme} élément diagonal de la matrice $\mathcal{V}$[LInd(T_2)]. Ou, en d'autres termes, le rapport des deux indices $\text{Ind}_t(T_2) / \text{Ind}_t(T_1)$ suit une loi log-normale $\mathcal{LN}(0 ; v(t))$.

Annexe 22 : Etude de la fonction f sur [0,T – 1]

On a pour i = 1, 2 :

$f_i(t) = (1 - \alpha_i^{T-t})\,(1 - \alpha_i^{t+1}) = [\,1 - \exp(-\lambda_i(T - t))\,]\,[\,1 - \exp(-\lambda_i(t+1))\,]$

$f_i(t) = [\,1 + \exp(-\lambda_i(T + 1))\,] - [\,\exp(-\lambda_i(T - t)) + \exp(-\lambda_i(t+1))\,]$

Posons $m = (T + 1)/2$, $a = (T – 1)/2 = m – 1$ et $u = t – a$; quand t parcourt [0, T – 1], la variable u parcourt l'intervalle [– a, a].

$f_i(t) = \exp(-\lambda_i m)\,[\,\exp(-\lambda_i m) + \exp(\lambda_i m) - \exp(\lambda_i u) - \exp(-\lambda_i u)\,]$

$f_i(t) = 2\exp(-\lambda_i m)\,[\,\cosh(\lambda_i m) - \cosh(\lambda_i u)\,]$ où $\cosh(x) = (e^{x} + e^{-x})/2$

On peut alors écrire :

$f(t) = \exp(-(\lambda_2 - \lambda_1)m)\,[\,\cosh(\lambda_2 m) - \cosh(\lambda_2 u)\,] / [\,\cosh(\lambda_1 m) - \cosh(\lambda_1 u)\,]$

Pour $\lambda_1 > \lambda_2$, la fonction f est croissante sur [0,a], elle atteint son maximum pour la valeur a, puis elle décroît. Elle est symétrique par rapport à l'axe vertical d'équation x = a. f est majorée par 1 et son maximum, noté $M(\lambda_1, \lambda_2, T)$, vaut :

$$\begin{aligned} M(\lambda_1, \lambda_2, T) &= \exp(-(\lambda_2 - \lambda_1)m)\,[\,\cosh(\lambda_2 m) - 1\,] / [\,\cosh(\lambda_1 m) - 1\,]. \\ &= [\,\exp(-\lambda_2 m)\,(\,\cosh(\lambda_2 m) - 1\,)\,] \,/\, [\,\exp(-\lambda_1 m)\,(\,\cosh(\lambda_1 m) - 1\,)] \\ &= [(\,\exp(-\lambda_2 m) - 1\,) \,/\, (\,\exp(-\lambda_1 m) - 1\,)\,]^2 \\ &= [\,d_2(m) / d_1(m)\,]^2 \end{aligned}$$

Annexe 23 : Illustration numérique de la formule de réversibilité

Le premier paragraphe de cette annexe calcule un indice BMN pour trois dates (t = 0, 1, 2) en utilisant deux méthodes : la relation matricielle $\hat{I}R = \eta P$ et l'estimation classique par moindres carrés. En prolongeant le temps à une quatrième date (t = 3), de nouvelles données de ventes répétées sont ajoutées à l'échantillon d'estimation et le deuxième paragraphe donne les nouvelles valeurs de l'indice pour cet échantillon global. On calcule ensuite l'indice que l'on obtient pour les 4 dates, en utilisant uniquement les nouvelles données. Puis, à partir de ce résultat, on vérifie dans le quatrième paragraphe que la formule de réversibilité est bien satisfaite. La dernière section reprendra ces calculs, mais en se plaçant cette fois dans un environnement de type Case-Shiller.

1. Calcul de l'indice pour 3 dates

1.1. Avec la formule $\hat{I}R = \eta P$

On considère dans un premier temps trois dates : t = 0, 1, 2

Données de base :

Date d'achat	Date de revente	Prix d'achat	Prix de revente	Durée de détention
0	1	101	108	1
0	1	102	110	1
0	2	99	122	2
0	2	104	124	2
1	2	110	118	1
1	2	112	109	1
1	2	114	125	1

Distribution des $n_{i,j}$

2	2
	3

Distribution des $L_{i,j}$

2	2
	3

On se place dans un contexte BMN où les tendances locales ne sont pas prises en compte. La quantité d'information d'une vente répétée ne dépend donc pas de la durée de détention. La notion d'information étant définie à une constante multiplicative près, on prendra alors la plus simple des conventions $L_{i,j} = n_{i,j}$

Matrice η

$$\begin{pmatrix} 4 & 0 \\ 0 & 5 \end{pmatrix}$$

Matrice $\hat{I}$

$$\begin{pmatrix} 4 & 2 \\ 2 & 5 \end{pmatrix}$$

Moyennes des prix d'achat : $h_p^{(i,j)}$

101,4988	101,4692
	111,9881

Moyennes des prix de vente : $h_f^{(i,j)}$

108,9954	122,9959
	117,1492

Prix moyens d'achat et de revente pour les investisseurs détenant de l'immobilier pendant l'intervalle [t,t+1] : $H_p(t)$ et $H_f(t)$

t	$H_p(t)$	$H_f(t)$
0	101,4840	115,7843
1	107,6556	119,4538

Système $\hat{I}R = \eta P$

Les coordonnées du vecteur ηP peuvent être obtenues sans passer par le calcul explicite de ρ_t et τ^t. Il suffit pour cela d'utiliser la formule adaptée au contexte BMN (cf. annexe 11) :

$$(n^t \zeta^t)\,\rho_t = (n^t \zeta^t) * (I^t / (n^t \zeta^t)) * \ln [H_f(t) / H_p(t)]$$
$$= I^t \ln [H_f(t) / H_p(t)]$$

Il vient :

$$\begin{cases} 4\,r_0 + 2\,r_1 = 0{,}52732 \\ 2\,r_0 + 5\,r_1 = 0{,}51995 \end{cases} \Leftrightarrow \begin{cases} r_0 = 9{,}98\,\% \\ r_1 = 6{,}41\,\% \end{cases}$$

Valeurs indicielles : $I_0 = 100$ $I_1 = 110{,}49$ $I_2 = 117{,}81$

1.2. Par la méthode des moindres carrés ordinaires

On reprend ici, pour l'estimation traditionnelle par moindres carrés, les notations utilisées dans Baroni et al. (2004)

Vecteur des rendements réalisés :

R = (0.06701 ; 0.07551 ; 0.20890 ; 0.17589 ; 0.07020 ; -0.02715 ; 0.09212)

Matrice D :

$$\begin{pmatrix} -1 & 0 \\ -1 & 0 \\ 0 & -1 \\ 0 & -1 \\ 1 & -1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix}$$

Equations normales : (D'D) LInd = D'R où LInd = (Ln(I_1/I_0) ; Ln(I_2/I_0)) et $I_0 = 100$

$$\begin{cases} 5\,\text{Ln}(I_1/I_0) - 3\,\text{Ln}(I_2) = 0{,}00735 \\ -3\,\text{Ln}(I_1/I_0) + 5\,\text{Ln}(I_2) = 0{,}51996 \end{cases} \Leftrightarrow \begin{cases} I_1 = 110{,}49 \\ I_2 = 117{,}81 \end{cases}$$

On retrouve évidemment les mêmes valeurs indicielles.

2. Calcul de l'indice prolongé à 4 dates

t prend toujours les valeurs 0, 1, 2 mais on prolonge le temps jusqu'à t = 3. Les nouvelles données sont présentées dans le tableau ci-dessous.

Date d'achat	Date de revente	Prix d'achat	Prix de revente	Durée de détention
0	1	101	110	1
0	2	102	120	2
1	2	105	122	1
1	3	112	130	2
2	3	122	132	1
2	3	115	136	1

En ajoutant ces nouvelles données aux anciennes on obtient les résultats suivants :

Distribution des $n_{i,j}$

3	3	0
	4	1
		2

Distribution des $L_{i,j}$

3	3	0
	4	1
		2

Matrice η

$$\begin{pmatrix} 6 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Matrice $\hat{I}$

$$\begin{pmatrix} 6 & 3 & 0 \\ 3 & 8 & 1 \\ 0 & 1 & 3 \end{pmatrix}$$

Moyennes des prix d'achat : $h_p^{(i,j)}$

101,3322	101,6458	**1**
	110,1986	112
		118,4483

Moyennes des prix de vente : $h_f^{(i,j)}$

109,3293	121,9891	**1**
	118,3435	130
		133,9851

Lorsqu'une classe (i,j) n'a pas d'éléments, nous fixons les valeurs de $h_p^{(i,j)}$ et $h_f^{(i,j)}$ à 1. Ce choix est imposé par les équations du modèle.

Prix moyens d'achat et de revente pour les investisseurs détenant de l'immobilier pendant l'intervalle [t,t+1] : $H_p(t)$ et $H_f(t)$

T	$H_p(t)$	$H_f(t)$
0	101,4889	115,4858
1	107,1270	121,1115
2	116,2586	132,6433

Système $\hat{I}R = \eta P$

$$\begin{cases} 6\,r_0 + 3\,r_1 + 0\,r_2 = 0{,}77519 \\ 3\,r_0 + 8\,r_1 + 1\,r_2 = 0{,}98158 \\ 0\,r_0 + 1\,r_1 + 3\,r_2 = 0{,}39554 \end{cases} \Leftrightarrow \begin{cases} r_0 = 9{,}17\ \% \\ r_1 = 7{,}49\ \% \\ r_2 = 10{,}69\ \% \end{cases}$$

Valeurs indicielles : $I_0 = 100$ $I_1 = 109,61$ $I_2 = 118,13$ $I_3 = 131,46$

Le phénomène de réversibilité est visible pour les valeurs I_1 et I_2 qui ont légèrement fluctué par rapport à la première estimation.

3. Indice construit uniquement avec les nouvelles données

Dans le paragraphe précédent l'indice a été calculé en utilisant les anciennes valeurs et les nouvelles valeurs. Cette partie présente les résultats que l'on obtient si l'on travaille uniquement avec les nouvelles données, rappelées ci-dessous :

Date d'achat	Date de revente	Prix d'achat	Prix de revente	Durée de détention
0	1	101	110	1
0	2	102	120	2
1	2	105	122	1
1	3	112	130	2
2	3	122	132	1
2	3	115	136	1

Distribution des $n_{i,j}$

1	1	0
	1	1
		2

Distribution des $L_{i,j}$

1	1	0
	1	1
		2

Matrice η

$$\begin{pmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix}$$

Matrice $\hat{I}$

$$\begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix}$$

Moyennes des prix d'achat : $h_p^{(i,j)}$

101	102	**1**
	105	112
		118,4483

Moyennes des prix de vente : $h_f^{(i,j)}$

110	120	**1**
	122	130
		133,9851

Prix moyens d'achat et de revente pour les investisseurs détenant de l'immobilier pendant l'intervalle [t,t+1] : $H_p(t)$ et $H_f(t)$

T	$H_p(t)$	$H_f(t)$
0	101,4988	114,8913
1	106,2517	123,9257
2	116,2586	132,6433

Système $\hat{I}R = \eta P$

$$\begin{cases} 2\,r_0 + 1\,r_1 + 0\,r_2 = 0{,}2479 \\ 1\,r_0 + 3\,r_1 + 1\,r_2 = 0{,}4616 \\ 0\,r_0 + 1\,r_1 + 2\,r_2 = 0{,}3955 \end{cases} \Leftrightarrow \begin{cases} r_0 = 7{,}64\ \% \\ r_1 = 9{,}50\ \% \\ r_2 = 10{,}02\ \% \end{cases}$$

Valeurs indicielles $I_0 = 100$ $I_1 = 107{,}94$ $I_2 = 118{,}70$ $I_3 = 131{,}21$

4. Vérification de la formule de réversibilité

En utilisant les notations du paragraphe 3, chapitre 3, avec $T_1 = 2$ et $T_2 = 3$, on a :

$R(T_1) = (0{,}0998 ; 0{,}0641 ; 0)'$ $\quad R(T_2 \backslash T_1) = (0{,}0764 ; 0{,}0950 ; 0{,}1002)'$

$R(T_2) = (0{,}0917 ; 0{,}0749 ; 0{,}1069)'$

Matrice $I(T_1)$ Matrice $I(T_2 \backslash T_1)$ Matrice $I(T_2)$

$$\begin{pmatrix} 4 & 2 & 0 \\ 2 & 5 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 2 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 3 \end{pmatrix} \quad \begin{pmatrix} 6 & 3 & 0 \\ 3 & 8 & 1 \\ 0 & 1 & 3 \end{pmatrix}$$

On peut alors vérifier numériquement que la relation de réversibilité est effectivement satisfaite :

$I(T_1)\, R(T_1) + I(T_2 \backslash T_1)\, R(T_2 \backslash T_1) = (0{,}7752 ; 0{,}9816 ; 0{,}3955)' = I(T_2)\, R(T_2)$

5. Vérification de la formule de réversibilité pour une situation Case-Shiller

A partir de ces mêmes données, si on suppose que le contexte est de type Case-Shiller (avec $\Theta = 10$), on peut vérifier que la formule de réversibilité est également satisfaite :

$R(T_1) = (0{,}0998 ; 0{,}0634 ; 0)'$ $\quad R(T_2 \backslash T_1) = (0{,}0761 ; 0{,}0964 ; 0{,}1010)'$

$R(T_2) = (0{,}0911 ; 0{,}0749 ; 0{,}1078)'$

Matrice $I(T_1)$ Matrice $I(T_2 \backslash T_1)$ Matrice $I(T_2)$

$$\begin{pmatrix} 0,35 & 0,17 & 0 \\ 0,17 & 0,44 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 0,17 & 0,08 & 0 \\ 0,08 & 0,26 & 0,08 \\ 0 & 0,08 & 0,27 \end{pmatrix} \quad \begin{pmatrix} 0,52 & 0,25 & 0 \\ 0,25 & 0,7 & 0,08 \\ 0 & 0,08 & 0,27 \end{pmatrix}$$

La vérification numérique s'écrit ici :

$I(T_1)\ R(T_1) + I(T_2 \backslash T_1)\ R(T_2 \backslash T_1) = (\ 0,0663\ ;\ 0,0840\ ;\ 0,0348\ )' = I(T_2)\ R(T_2)$

Annexe 24 : Echantillon de ventes répétées servant de support aux calculs des indices informationnels

Date d'achat	Date de revente	Prix d'achat	Prix de revente	Durée de detention
0	3	101	108	3
0	1	102	110	1
0	2	99	105	2
0	2	104	102	2
1	2	110	107	1
1	3	112	120	2
1	2	114	98	1
0	1	101	110	1
0	2	102	95	2
1	2	105	122	1
1	3	112	115	2
2	3	79	100	1
0	3	89	60	3

Annexe 25 : Détermination de δ_{inf} et δ_{sup}

La fonction tangente étant strictement croissante on peut raisonner indifféremment avec δ_i ou avec $tg(\delta_i)$. Les valeurs extrêmes pour δ_i et $tg(\delta_i)$ seront en effet atteintes pour les mêmes coordonnées du vecteur v_i.

δ_i est un angle géométrique (positif) et non pas algébrique, on a donc nécessairement $\delta_{inf} \geq 0$. La valeur 0 peut être atteinte par δ_i, ou de manière équivalente par $tg(\delta_i)$, quand :

$$E(v^2_i) = [E(v_i)]^2 \Leftrightarrow \sigma(v_i) = 0 \Leftrightarrow v_i(j) = v_i(i) \quad \text{pour tous les j}$$

On a donc $\delta_{inf} = 0$ et δ_i atteint ce niveau quand v_i est colinéaire à d^+, c'est-à-dire quand $v_i = v'_i$.

La borne supérieure est moins immédiate à déterminer, toutefois on peut l'obtenir en écrivant :

$$E(v^2_i) / [E(v_i)]^2 = [\, E(v^2_i) / E(v_i) \,] / E(v_i)$$
$$= [(v^2_i(1) + \ldots + v^2_i(T)) / (v_i(1) + \ldots + v_i(T))] / E(v_i)$$

Comme $v_i(j) \leq v_i(i)$, on a aussi $v^2_i(j) \leq v_i(i)\, v_i(j)$ et en sommant ces inégalités on arrive à :

$$E(v^2_i) / [E(v_i)]^2 \leq v_i(i) / E(v_i) \leq T\, E(v_i) / E(v_i) \leq T$$

On a ainsi $\delta_{sup} \leq tg^{-1}((T-1)^{1/2})$. Il est possible d'atteindre cette borne si et seulement si les deux conditions suivantes sont satisfaites :

$$v^2_i(1) + \ldots + v^2_i(T) = v_i(i)\, (v_i(1) + \ldots + v_i(T)) \text{ et } \quad v_i(i) = T\, E(v_i)$$

La deuxième condition donne directement $v_i(j) = 0$ pour tous les $j \neq i$, satisfaisant ainsi simultanément la première condition. La borne supérieure $\delta_{sup} = tg^{-1}((T-1)^{1/2})$ correspond donc à la situation où v_i est sur son axe (Ox_i).

Annexe 26 : Modification marginale de l'information fournie par v_i

En utilisant les notations suivantes : $a = \sum_{k\neq j} v_i(k)$ et $b = \sum_{k\neq j} v^2_i(k)$

on peut écrire : $\text{Cos}\,(\delta_i) = (v_i(j) + \sum_{k\neq j} v_i(k)) / [T (v^2_i(j) + \sum_{k\neq j} v^2_i(k))]^{1/2}$

En dérivant par rapport à $x = v_i(j)$ on obtient :

$$T [\sum_{k\neq j} v^2_i(k) - x \sum_{k\neq j} v_i(k)] / [T (x^2 + \sum_{k\neq j} v^2_i(k))]^{3/2}$$

Cette expression est positive si et seulement si :

$$\begin{aligned} v_i(j) \sum_{k\neq j} v_i(k) < \sum_{k\neq j} v^2_i(k) &\Leftrightarrow v_i(j) \sum_k v_i(k) < \sum_k v^2_i(k) \\ &\Leftrightarrow v_i(j) < \left(\sum_k v^2_i(k) \,/\, \sum_k v_i(k) \right) \\ &\Leftrightarrow v_i(j) < E(v^2_i) / E(v_i) \\ &\Leftrightarrow v_i(j) < [V(v_i) + [E(v_i)]^2] / E(v_i) \\ &\Leftrightarrow v_i(j) < E(v_i) [1 + (\sigma(v_i)/E(v_i))^2] \\ &\Leftrightarrow v_i(j) < E(v_i) [1 + tg^2 (\delta_i)] \end{aligned}$$

En notant $\text{diff}(v_i) = E(v_i) [1 + tg^2 (\delta_i)]$ et en utilisant la décroissance de la fonction cosinus sur l'intervalle [0 ; 90], on établit donc que la fonction δ_i , de variable $v_i(j)$, décroît sur $[0, \text{diff}(v_i)]$ et croit au-delà de $\text{diff}(v_i)$.

Bibliographie

Ambrose, Buttimer. 2000. “Embedded options in the mortgage contract”. *Journal of real estate finance and economics* 21(2) : 95-111

Ambrose, Buttimer, Thibodeau. 2001. “A new spin on the Jumbo/Conforming loan rate differential”. *Journal of real estate finance and economics* 23(3) : 309-335

ANIL. 1999. "Expertise et negative equity". *Habitat actualités* Mai 1999

Azevedo-Pereira, Newton, Paxson. 2000. “Numerical solution of a two state variable contingent claims mortgage valuation model”. *Portuguese review of financial markets* 3 : 35-65

Azevedo-Pereira, Newton, Paxson. 2002. “UK fixed rate repayment mortgage and mortgage indemnity valuation”. *Real estate economics* 30(2) : 185-211

Azevedo-Pereira, Newton, Paxson. 2003. “Fixed-rate endowment mortgage and mortgage indemnity valaution”. *Journal of real estate finance and economics* : 26(2/3) : 197-221

Bailey, Muth, Nourse. 1963. “A regression method for real estate price index construction”. *Journal of the American Statistical Association* Vol 58

Bao, Wan. 2004. "On the use of spline smoothing in estimating hedonic housing price models: empirical evidence using Hong-Kong data". *Real estate economics* 32(3) : 487-507

Baroni, Barthélémy, Mokrane. 2004. “Physical real estate : A Paris repeat sales residential index”. *ESSEC Working paper* DR 04007, ESSEC Research Center, ESSEC Business School

Batsch. 2005. "Le financement adossé de l'immobilier en gestion de patrimoine : une modélisation simple". *Cahiers de recherche du CEREG* 2005-1

Baude, Bosvieux. 2002. "Hypothèque ou caution : l'exception française". *Habitat actualités* Décembre 2002

Beauvois, David et al. (2005). “Les indices Notaires INSEE de prix des logements anciens” *INSEE Méthode*, N° 111.

Berg. 2005. "Price indexes for multi-dwelling properties in Sweden" *Journal of real estate research* 27(1) : 47-81

Björk. (1998). “Arbitrage theory in continuous time”. *Oxford University Press*

Björk, Clapham. 2002. "A note on the pricing of real estate index linked swaps" *SSE / EFI working paper series in economics and finance,* N° 492

Black, Scholes. 1973. “The pricing of options and corporate liabilities”. *Journal of political economy* 81 : 637-654

Bluhm, Overbeck, Wagner. 2003. "An introduction to Credit Risk modeling". *Chapman & Hall/CRC Financial mathematics series.*

Bosvieux, Vorms. 2003. "Durée des prêts : allongement conjoncturel ou changement d'attitude à l'égard de l'endettement? " *Habitats actualités* Mai 2003

Brown, Matysiak. 2000. "Sticky valuations, aggregation effects, and property indices". *Journal of real estate finance and economics* 20(1) : 49-66

Buttimer, Kau, Slawson. 1997. "A model for pricing securities dependent upon a real estate index" *Journal of housing economics* 6 : 16-30

Caisse centrale du crédit immobilier de France, 3CIF. Comptes consolidés et rapport de gestion 2003. *Document de référence déposé à l'AMF le 22/07/04*

Cannaday, Munneke, Yang. 2005. "A multivariate repeat-sales model for estimating house price indices" *Journal of urban economics* 57(2) : 320-342

Case, Pollakowski, Watcher. 1991. "On choosing among house price index methodologies" *AREUEA Journal* 19(3) : 286-307

Case, Quigley. 1991. "The dynamics of real estate prices". *The review of economics and statistics* 73 (1) : 50-58

Case, Shiller. 1987. "Prices of single family homes since 1970: new indexes for four cities". *New England Economic Review* September/October 1987 : 45-56.

Case, Shiller. 1989. "The efficiency of the market for single-family homes". *The American economic review* 79(1) : 125-137

Chau, Wong, Yiu. 2005. "Adjusting for non-linear age effects in the repeat sales index" *The journal of real estate finance and economics* 31(2) : 137-153

Chau, Wong, Yiu, Leung. 2005. "Real estate price indices in Hong-Kong" *Journal of real estate literature* 13(3) : 337-356

Cheung, Yau, Hui. 2004. "The effects of attributes on the repeat sales pattern of residential property in Hong-Kong" *Journal of real estate finance and economics* 29(3) : 321-339

Childs, Ott, Riddiough. 2001. “Valuation and information acquisition policy for claims written on noisy real assets”. *Financial Management* summer 2001 : 45-75

Childs, Ott, Riddiough. 2002a. “Optimal valuation of noisy real assets”. *Real estate economics* 30(3) : 385-414

Childs, Ott, Riddiough. 2002b. “Optimal valuation of claims on noisy real assets: theory and application”. *Real estate economics* 30(3) : 415-443

Childs, Ott, Riddiough. 2004. “Effects of noise on optimal exercise decisions: the case of risky debt secured by renewable lease income”. *Journal of real estate finance and economics* 28(2/3) : 109-121

CIF Euromortgage. Comptes sociaux au 31/12/2003. *Disponible à www.cifeuromortgage.com*

CIFD. Comptes consolidés du groupe Crédit Immobilier de France. 2003. *Disponible à www.cifeuromortgage.com*

Clapp, Giaccotto. 1992. "Estimating price indices for residential property : a comparison of repeat sales and assessed value methods" *Journal of the American statistical association* 87(418) : 300-306

Clapp, Giaccotto. 1998. "Residential hedonic models : a rational expectations approach to age effects" *Journal of urban economics* 44(3) : 415-437

Clapp, Giaccotto. 1999. “Revisions in repeat-sales price indexes: Here today, gone tomorrow? ” *Real estate economics* 27(1) : 79-104

Clapp. 2004. "A semiparametric method for estimating local house price indices" *Real estate economics* 32(1) : 127-160

Compagnie de financement foncier. Rapport annuel 2004. *Document de référence déposé à l'AMF le 08/04/05*

Court. 1939. "Hedonic price indexes with automotive examples" In: *The dynamics of Automobile demand, General motors, New-York*

Cox, Oakes. (1984). "Analysis of Survival Data". London: Chapman and Hall.

CRH, Caisse de refinancement de l'habitat. Rapport annuel 2004. *Document de référence déposé à l'AMF le 07/02/05*

David, Dubujet, Gourriéroux, Laferrère. 2002. "Les indices de prix des logements anciens". *INSEE Méthode*, N° 98.

Demey, Frachot, Riboulet. 2003. "Introduction à la gestion actif-passif bancaire". *Economica*

Deng, Quigley, Van Order. 2000. "Mortgage terminations, heterogeneity and the exercise of mortgage options". *Econometrica* 68(2) : 275-308

Deville, Riva. 2004. "The determinants of the time to efficiency in options markets : a survival analysis approach". *Working paper* JEL classification : C41, G13, G14

Dombrow, Turnbull. 2004. "Trends in real estate research, 1988-2001 : What's hot and what's not". *Journal of real estate finance and economics* 29(1) : 47-70

Downing, Stanton, Wallace. 2003. “An empirical test of a two-factor mortgage valuation model : How much do house prices matter?”. *Working paper*

Dreiman, Pennington-Cross. 2004. "Alternative methods of increasing the precision of weighted repeat sales house prices indices" *Journal of real estate finance and economics* 28(4) : 299-317

Dunn, McConnell. 1981. “Valuation of GNMA Mortgage-Backed Securities”. *Journal of finance* 36(3) : 599-616

Duon. 1943a. "Evolution de la valeur vénale des immeubles parisiens". *Journal de la Société de Statistique de Paris* octobre.

Duon. 1943b."Evolution de la valeur vénale des immeubles à Paris de 1840 à 1939". *Bulletin Statistique de la France* décembre.

Duon. 1946. "Document sur le problème du logement". *Etudes Economiques* 1.

Eichholtz. 1997. "A long run house price index : the Herengracht index, 1628-1973" *Real estate economics* 25(2) : 175-192

Eitrheim, Erlandsen. 2004. "House prices in Norway 1819-1989" *Working paper*

Engberg, Greenbaum. 1999. "State enterprise zones and local housing markets". *Journal of housing research* 10(2) : 163-187

Englund, Quigley, Redfearn. 1998. "Improved price indexes for real estate : measuring the course of Swedish housing prices" *Journal of urban economics* 44(2) : 171-196

Englund, Hwang, Quigley. 2002. "Hedging housing risk". *Journal of real estate finance and economics* 24(1/2) : 167-200

Evans, Kolbe. 2005. "Homeowners' repeat-sale gains, dual agency and repeated use of the same agent" *Journal of real estate research* 27(3) : 267-292

Favereau, Biencourt, Eymard-Duvernay. (2002). "Conventions and structures in economic organization : markets, hierarchies and networks" *Edward Elgar, Cheltenham* Chap 8 : 213-252

Fisher, Gatzlaff, Geltner, Haurin. 2003. "Controlling for the impact of variable liquidity in commercial real estate price indices" *Real estate economics* 31(2): 269-303

Fisher, Geltner, Pollakowski. 2005. “A quarterly transactions-based index of institutional real estate investment performance and movements in supply and demand”. *Working paper*

Flavin, Yamashita. 2002. "Owner-occupied housing and the composition of the household portfolio" *The American economic review* 92(1) : 345-362

Francke, Vos. 2004. "The hierarchical trend model for property valuation and local price indices". *Journal of real estate finance and economics* 28(2/3) : 179-208

Frachot, Gouriéroux. 1995. "Titrisation et remboursements anticipés". *Economica*

Friggit. 2001. "Prix des logements, produits financiers immobiliers et gestion des risques". *Economica*

Friggit. 2002. "Placement en actions et en logement : quelques régularités sur longue période". *Réfexions immobilières* 33

Gatzlaff, Geltner. 1998. "A repeat-sales transaction-based index of commercial property" *A study for the real estate research institute*

Gatzlaff, Haurin. 1998. "Sample selection and biases in local house value indices" *Journal of urban economics* 43(2) : 199-222

Geltner. 1991. "Smoothing and appraisal-based returns" *Journal of real estate finance and economics* 4(3) : 327-345

Geltner, Goetzmann. 2000. "Two decades of commercial property returns : a repeated-measures regression-based version of the NCREIF index" *Journal of real estate finance and economics* 21(1) : 5-21

Girardin, Limnios. 2001. "Probabilités" *Vuibert*, collection "Les grand cours"

Goetzmann. 1992. “The accuracy of real estate indices: repeat sales estimators”. *Journal of real estate finance and economics* 5 : 5-53

Goetzmann. 1993. "Accounting for taste : art and the financial markets over three centuries" *American economic review* 83 : 1370-1376

Goetzmann, Peng. 2002. "The bias of the RSR estimator and the accuracy of some alternatives" *Real estate economics* 30(1) : 13-39

Heisenberg. (1932). “Les principes physiques de la théorie des quanta”. Réédition par Jacques Gabay (1989) ISBN 2-87647-080-2

Heston. 1993. “A closed-form solution for options with stochastic volatility with applications to bond and currency options”. *The review of financial studies* 6(2) : 327-343

Hill, Sirmans, Knight. 1999. "A random walk down main street? " *Regional science and urban economics* 29(1) : 89-103

Hilliard, Kau, Slawson. 1998. “Valuing prepayment and default in a fixed-rate mortgage : a bivariate binomial options pricing technique”. *Real estate economics* 26(3) : 431-468

Hoesli, Giaccotto, Favarger. 1997. "Three new real estate price indices for Geneva, Switzerland". *Journal of real estate finance and economics* 15(1) : 93-109

Hoesli, Lekander, Witkiewicz. 2004. "International evidence on real estate as a portfolio diversifier" *Journal of real estate research* 26(2) : 161-206

Hoesli, Thion, Watkins. 1997. "A hedonic investigation of the rental value of apartments in central Bordeaux". *Journal of property research* 14(1) : 15-26

Hordijk, De Kroon, Theebe. 2004. "Long-run return series for the European continent : 25 years of Dutch commercial real estate" *Journal of real estate portfolio management* 10(3) : 217-230

Iacoviello, Ortalo-Magné. 2003. "Hedging housing risk in London". *Journal of real estate finance and economics* 27(2) : 191-209

Jachiet, Friggit, Vorms, Taffin. 2004. “Rapport sur le prêt viager hypothécaire et la mobilisation de l’actif résidentiel des personnes âgées”. *Inspection générale des*

finances, Conseil général des Ponts et Chaussées, ANIL, available at : http://www.anil.org/index.htm

Kaiser. 2004. "Real estate as a surrogate for bonds : a dynamic asset allocation view" *Journal of real estate portfolio management* 10(1) : 23-35

Kalbfleisch, Prentice. 2002. “The Statistical Analysis of Failure Time Data” *Wiley-Interscience*

Kau, Keenan, Muller, Epperson. 1992. “A generalized valuation model for fixed-rate residential mortgages”. *Journal of money, credit and banking* 24(3) : 279-299

Kau, Keenan, Muller, Epperson. 1993. “Option theory and floating-rate securities with a comparison of adjustable and fixed-rate mortgages” *Journal of business* 66(4) : 595-618

Kau, Keenann. 1995. “An overview of the option-theoretic pricing of mortages”. *Journal of housing research* 6(2) : 217-244

Kau, Slawson. 2002. “Frictions, heterogeneity and optimality in mortgage modeling”. *Journal of real estate finance and economics* 24(3) : 239-260

Kelly, Slawson. 2001. “Time-varying mortgage prepayment penalties”. *Journal of real estate finance and economics* 23(2) : 235-254

Kindleberger. 2004. “Histoire mondiale de la spéculation financière” *Valor Editions*

Kleiman, Payne, Sahu. 2002. "Random walks and market efficiency : evidence from international real estate markets" *Journal of real estate research* 24(3) : 279-297

Knight, Dombrow, Sirmans. 1995. "A varying parameters approach to constructing house price indexes" *Real estate economics* 23(2) : 187-205

Lacour-Little, Malpezzi. 2003. "Appraisal quality and residential mortgage default : evidence from Alaska" *Journal of real estate finance and economics* 27(2) : 211-233

Lai, Wang. 1998. "Appraisal smoothing : the other side of the story" *Real estate economics* 26(3) : 511-535

Laferrère. (2004). "Hedonic housing price indexes: the French experience" *Working paper*

Lautier. 2002a. “Les options réelles : Une idée séduisante – Un concept utile et multiforme – Un instrument facile à créer mais difficile à valoriser”. *Working paper, Cahiers de recherche du CEREG 2002-05, Université Paris IX, www.dauphine.fr/cereg/listecahiers.htm*

Lautier. 2002b. “Il y a des loups dans la forêt des options réelles” *Working paper, Cahiers de recherche du CEREG 2002-06, Université ParisIX, www.dauphine.fr/cereg/listecahiers.htm*

Le Blanc, Lagarenne. 2004. " Owner-occupied housing and the composition of the household portfolio : the case of France" *Journal of real estate finance and economics* 29(3) : 259-275

Lee. 2005. "The return due to diversification of real estate to the US mixed-asset portfolio" *Journal of real estate portfolio management* 11(1) : 19-28

Lee, Stevenson. 2005. "The case for REITs in the mixed-asset portfolio in the short and long run". *Journal of real estate portfolio management* 11(1) : 55-80

Leung, Leong, Wong. 2006. "Housing price dispersion : an empirical investigation" *Journal of real estate finance and economics* 32(3) : 357-385

Mahé de Boislandelle. 2005. "Marché de l'art et gestion de patrimoine" *Economica*

Marcato. 2004. "Style analysis in real estate markets and the construction of value and growth indexes" *Journal of real estate portfolio management* 10(3) : 203-215

Markowitz. 1952. “Portfolio selection” *Journal of finance* 7(1) : 77-91

McConnell, Singh. 1994. “Rational prepayments and the valuation of collateralized mortgage obligations”. *Journal of finance* 49(3) : 891-921

McMillen, Dombrow. 2001. "A flexible Fourier approach to repeat sales price indexes" *Real estate economics* 29(2) : 207-225

Meese, Wallace. 1997. “The construction of residential housing price indices: A comparison of repeat-sales, hedonic-regression and hybrid approaches”. *Journal of real estate finance and economics* 14 : 51-73

Merton. 1973. “Theory of rational option pricing”. *Bell journal of economics and management Science* 4 : 141-183

Merton. 1971. “Optimum consumption and portfolio rules in a continuous time model”. *Journal of economic theory* 3 : 373-413

Morieux. 1980. "Satisfaction, profit et risqué retirés d'un placement en actifs non financiers" *Thèse en sciences de gestion.* Université de Lille 1

Munneke, Slade. 2000. "An empirical study of sample-selection bias in indices of commercial real estate" *Journal of real estate finance and economics* 21(1) : 45-64

Munneke, Slade. 2001. "A metropolitan transaction-based commercial price index : a time-varying parameter approach". *Real estate economics* 29(1) : 55-84

Nappi-Choulet, Maleyre, Maury. 2005. "A hedonic price model for the office market: an application for the Paris-La Défense district" *Communication présentée à la conférence 2005 de l'ERES à Dublin*

Newell, Worzala, McAllister, Schulte. 2004. "An international perspective on real estate research priorities" *Journal of real estate portfolio management* 10(3) : 161-170

Ong, Lusht, Mak. 2005. "Factors influencing auction outcomes : bidder turnout, auction houses and market conditions" *Journal of real estate research* 177-191

Peng. 2002. "GMM repeat sales price indices" *Real estate economics* 30(2) : 239-261

Philippe. 2004. “New Boundaries to Real Options Valuation? Exploratory Research Based on a Case Study”. *Working paper, Cahiers de recherche du CEREG 2004-03, Université Paris IX, www.dauphine.fr/cereg/listecahiers.htm*

Quigley. 1995. “A simple hybrid model for estimating real estate price indexes”. *Journal of housing economics* 4 : 1-12

Rosen. 1974. “Hedonic price and implicit markets : product differentiation in pure competition” *Journal of political economy* 1

Schoeffler. 2005. "Foncières cotées, SCPI, actions et obligations : les enseignements de l'allocation d'actifs". *Réflexions immobilières* 40 : 19-26

Schwartz, Torous. 1989. “Prepayment and the valuation of Mortgage-Backed Securities”. *Journal of finance* 44(2) : 375-392

Shiller. 1991. “Arithmetic repeat sales price estimators” *Journal of housing economics* 1 : 110-126

Simon. 2004. "Modélisation des mortgages, une revue de la littérature". *Mémoire de DEA, prix ASF 2004.*

Stanton. 1995. “Rational prepayment and the valuation of mortgage-backed securities”. *Review of financial studies* 8(3) : 677-708

Stevenson. 2000. "International real estate diversification : empirical tests using hedged indices" *Journal of real estate research* 19(1/2) : 105-131

Sunderman, Spahr, Birch, Oster. 2000. "Impact of ranch and market factors on an index of agricultural holding period returns" *Journal of real estate research* 19(1/2) : 209-234

Thion. 1998. "Valeur, prix et méthodes d'évaluation en immobilier" *Cahiers de recherche du CEREFI* 14(98)

Titman, Torous. 1989. “Valuing commercial mortgages: an empirical investigation of the contingent-claims approach to pricing risky debt”. *Journal of finance* 44(2) : 345-373

Titrilog 11-98. 1998. “Note d’information”. *ABC Gestion – Crédit Lyonnais*

Tu, Yu, Sun. 2004. "Transaction-based office price indexes : a spatiotemporal modelling approach". *Real estate economics* 32(2) : 297-328

Tong, Glascock. 2000. "Price dynamics of owner-occupied housing in the Baltimore-Washington area : Does structure type matter ?". *Journal of housing research* 11(1) : 29-66

Vauban Mobilisations Garanties. Rapport annuel 2003. *Document de référence déposé à l’AMF le 08/10/04*

Wang, Zorn. 1997. “Estimating house price growth with repeat sales data: What’s the aim of the game?” *Journal of housing economics* 6 : 93-118

Wolverton, Senteza. 2000. "Hedonic estimates of regional constant quality house prices" *Journal of real estate research* 19(3) : 235-253

Young. 2005. "Making sense of the NCREIF property index : a new formulation revisited" *Journal of real estate portfolio management* 11(3) : 211-223

Vu : le Président

M……..

Vu : les suffragants

MM…….

Vu et permis d'imprimer :

Le Vice-Président du Conseil Scientifique Chargé de la Recherche de l'Université Paris-Dauphine.

RESUME

L'immobilier se financiarise. L'époque où ce secteur n'était pas considéré comme une classe d'actifs à part entière est révolue. Mais une fois que l'on a pris acte de cet état de fait, la question est de savoir comment travailler avec cet actif récalcitrant, hétérogène et illiquide. Les indices immobiliers fournissent des éléments de réponse ; notamment pour la gestion de portefeuille, la gestion des risques avec les produits dérivés et la gestion des prêts immobiliers. Dans cette thèse nous développons et nous approfondissons la structure de l'indice de ventes répétées de Case et Shiller en la rendant plus interprétable financièrement, plus maniable et plus intuitive. Nous étudions ainsi le lien fonctionnel entre les indices de prix et l'indice de ventes répétées. Nous présentons une méthodologie d'analyse de données qui, grâce à divers indicateurs, assure une exploitation de l'information enchâssée dans les échantillons bien supérieure à celle que l'on obtient en se contentant d'utiliser la procédure traditionnelle. Nous étudions la fiabilité de l'indice, sa volatilité et le problème des deux populations. Un procédé de quantification du phénomène de réversibilité est aussi développé et des éléments pour l'évaluation des produits dérivés sur indices immobiliers sont présentés. Enfin, cette problématique aboutira à l'introduction du concept d'indice informationnel et l'on posera la question centrale pour toute construction d'indice : Comment quantifier l'information ?

Mots clés : Immobilier, Indices Immobiliers, Ventes répétées, Information, Réversibilité

ABSTRACT

Real estate is becoming more and more a financial asset. Nowadays it fully appears as an asset class like stocks or bonds. But once this diagnosis is layed down, how can we work with this illiquid, heterogenous and recalcitrant asset ? Real estate indexes bring some elements, especially for portfolio management, risk management and mortgages. In this dissertation we redevelop and deepen the repeat-sales framework of Case and Shiller, making it easier to interpret financially, easier to handle and more intuitive. We study the theoretical link between the price indexes and the repeat-sales index (RSI). We present a methodology of data analysis which, thanks to various indicators, allows a better exploitation of the information embedded in a dataset, compared to the computation of the single index values usually realised. We also study the sensibility, the volatility of the RSI and the problem of the two populations. We develop a quantification process of the reversibility phenomenon and discuss the possibility to price derivatives written on a real estate index. Finally, this work ends up with the introduction of the concept of informational index, where the crucial question for every index computation in a heterogeneous and illiquid market is formulated : How can information be quantified ?

Key words : Real Estate, Real Estate Indexes, Repeat-sales, Information, Reversibility