

DOCUMENT D'APPUI No 5 (16)*

24 octobre 2001

FRANÇAIS SEULEMENT

COMMISSION DE STATISTIQUE et
COMMISSION ÉCONOMIQUE
POUR L'EUROPE

ORGANISATION INTERNATIONALE du
TRAVAIL (OIT)

CONFÉRENCE DES STATISTICIEANS
EUROPÉENS

Réunion commune CEE/OIT sur les indices
Des prix à la consommation
(Genève, 1-2 novembre 2001)

EVALUATION DES MICRO-DONNEES ET DE L'IPC
CAS DES LAVE-VAISSELLE ET DES TELEVISIONS

Document d'appui présenté par INSEE, Direction des Statistiques Démographiques et
Sociales, France**

? Document placé sur Internet tel que soumis par le pays.

** Préparé par Mme. Isabelle LE BOETTE et M. Christophe BARRET, INSEE.

Septembre 2001

EVALUATION DES MICRO-DONNEES ET DE L'IPC

CAS DES LAVE-VAISSELLE ET DES TELEVISIONS

CONFERENCE ONU / OIT
SUR LES INDICES DE PRIX A LA CONSOMMATION
1 et 2 novembre 2001
Genève

*Isabelle LE BOETTE
Christophe BARRET*

*INSEE, Direction des Statistiques Démographiques et Sociales
Division Prix à la Consommation, Timbre F320
18, Boulevard Adolphe Pinard, 75675 PARIS CEDEX 14, FRANCE
E-mail : isabelle.le-boette@insee.fr
christophe.barret@insee.fr*

RESUME

Cet article se propose de mesurer les bénéfices que peut tirer l'Indice des Prix à la Consommation (IPC) de l'utilisation de micro-données. Il s'appuie pour cela sur l'exemple de deux biens durables, les lave-vaisselle et les téléviseurs.

Un premier constat est le fort coût d'entrée : de nombreux traitements sont nécessaires avant de disposer de fichiers satisfaisant aux critères de cohérence des caractérisations techniques et des prix. Le travail occasionné est tel qu'actuellement il réduit la faisabilité d'utiliser des micro-données directement en production. On peut cependant mettre à profit le fait que ces données fournissent une estimation de l'exhaustivité du marché. Elles peuvent alors aider à l'amélioration de la représentativité des échantillons, en constituant une référence à laquelle on compare les caractéristiques du champ de collecte IPC. Les techniques de comparaison présentées s'appuient d'abord sur la constitution de fichiers GFK et IPC aux mêmes niveaux d'agrégation, puis sur l'examen des prix moyens et des chiffres d'affaires déclinés selon plusieurs critères. Les résultats obtenus sont enrichissants, puisqu'ils mettent en évidence les points forts et les lacunes de notre processus de sélection des produits en vue du calcul de l'IPC. Enfin, quelques premières comparaisons d'indices sont présentées, qui laissent apparaître entre-autres des différences de prise en compte des substitutions par nos deux sources de données.

SOMMAIRE

INTRODUCTION	5
1. TRAITEMENT DES MICRO-DONNEES GFK.....	6
1.1. METHODOLOGIE GFK.....	6
1.1.1. Plan de sondage et estimation.....	6
1.1.2. Suivi de l'échantillon dans le temps.....	7
1.1.3. Enquêtes.....	7
1.1.4. Disponibilité des données.....	7
1.2. DIFFICULTES DE MISE EN FORME DES DONNEES.....	7
1.2.1. Description des fichiers.....	7
1.2.2. Uniformisation des fichiers Excel en vue d'une importation sous SAS.....	8
1.2.3. Création d'un fichier par produit.....	8
1.3. CARACTERISATION DES DONNEES.....	9
1.3.1. Présentation.....	9
1.3.2. Détection des problèmes.....	9
1.3.3. Traitement.....	11
1.3.4. Agrégation.....	14
1.3.5. Evaluation.....	14
1.4. TRAITEMENT DES INCOHERENCES SUR LES PRIX.....	15
1.4.1. Repérage et traitement des prix aberrants.....	15
1.4.2. Le choix des seuils.....	17
1.4.3. Impact des traitements réalisés.....	18
2. LES DONNEES IPC.....	20
2.1. PRESENTATION DES BIENS DURABLES DANS L'INDICE DES PRIX A LA CONSOMMATION.....	20
2.1.1. Un secteur spécifique.....	20
2.1.2. Formule de calcul d'un indice de variété hétérogène.....	20
2.2. TRAITEMENT DES DONNEES IPC.....	21
2.2.1. Rapatriement des données.....	21
2.2.2. Jointure entre l'IPC et GFK.....	21
2.2.3. Agrégation des données IPC.....	21
2.2.4. Calcul d'indices comparés.....	23
3. EVALUATION DE L'ECHANTILLON IPC A L'AIDE DES DONNEES GFK	24
3.1. DONNEES DE CADRAGE.....	24
3.1.1. Pondérations.....	24
3.1.2. Prix moyens.....	25
3.2. REPARTITION PAR FORME DE VENTE.....	28
3.3. MARQUE.....	29
3.4. MODELE.....	31
3.5. CARACTERISTIQUES TECHNIQUES.....	33
3.5.1. Cas des lave-vaisselle.....	33
3.5.2. Téléviseurs : taille d'écran.....	33
3.5.3. Téléviseurs : format.....	34
3.6. INDICES.....	34
CONCLUSION	37
ANNEXE : DICTIONNAIRE DES DONNEES	38
N°1 : LISTE DES VARIABLES GFK.....	38
N°2 : LISTE DES VARIABLES RETENUES.....	40
BIBLIOGRAPHIE.....	42

INTRODUCTION

L'intérêt des constructeurs et distributeurs pour les micro-données est aisément compréhensible. L'information relative à la taille, à la structure et aux parts de marché leur permet de se situer dans l'espace concurrentiel, et de définir en conséquence leur politique de production et de commercialisation. Depuis peu, les offices nationaux de statistiques trouvent eux aussi de plus en plus d'attrait à ce type de données. Au rang des bénéficiaires possibles, figurent les indices de prix. Ces indicateurs conjoncturels sont en effet des candidats permanents à l'amélioration, du fait de la perpétuelle contradiction entre les exigences de panier constant et d'échantillon représentatif. Les Biens Durables en particulier constituent un terrain propice à l'étude : le rythme soutenu des évolutions technologiques pose, plus que tout ailleurs, le problème du traitement de l'effet qualité et de la substitution entre produits.

Cet article fait suite à l'acquisition par l'INSEE de micro-données concernant quelques biens durables (téléviseurs, lave-vaisselle, micro-ordinateurs et imprimantes) auprès de la société GFK, spécialisée dans les biens d'équipement et de loisirs. Son objectif est d'évaluer les bénéfices que l'on peut tirer des micro-données pour le calcul de l'Indice des Prix à la Consommation (IPC), à partir de l'étude des téléviseurs et des lave-vaisselle.

Pour ce faire, une phase préalable de traitement des micro-données s'avère indispensable. Elle consiste en un travail d'uniformisation et de correction non négligeable. Il faut ensuite intervenir sur les données IPC, afin de leur attribuer une structure analogue : il s'agit notamment de passer à un niveau d'agrégation supérieur, occultant l'aspect géographique ; pondérations et prix moyens doivent alors être définis. Après quoi il devient possible de créer des outils d'évaluation de la qualité de notre indice, en se servant des données GFK comme d'une représentation de l'exhaustivité du marché. Cette évaluation repose principalement sur l'examen de la répartition, suivant différents critères, des chiffres d'affaires et des prix moyens. Elle est complétée par une première comparaison d'indices, donnant lieu à plusieurs hypothèses d'interprétation.

1. TRAITEMENT DES MICRO-DONNEES GFK

1.1. Méthodologie GFK

L'INSEE a acquis des micro-données sur biens durables auprès du département électronique grand public de la société GFK Marketing Services France. Avant d'étudier en détail cette source de données, voici les grands principes de sa méthodologie d'enquête.

1.1.1. Plan de sondage et estimation

En premier lieu, GFK réalise une stratification de l'univers des points de vente selon deux critères : la région géographique et le circuit de distribution. Les DOM et la Corse ne sont pas enquêtés. Le champ géographique de l'enquête est donc la France métropolitaine. GFK définit 4 types de circuits de distribution : les hypermarchés, les grands magasins spécialisés, les magasins traditionnels et les autres formes de ventes (grands magasins, vente par correspondance, ...). Les ventes sur Internet sont depuis peu intégrées au champ de l'enquête. Pour cela, GFK distingue deux catégories de sites : les « click-and-mortar » qui sont des sites de magasins (tel que la FNAC) et les « pure players » qui sont des sites commerciaux directs (tel que le site AMAZON). Les « click-and-mortar » sont classés dans la même forme de vente de l'enseigne et les « pure players » sont rangés dans les autres formes de vente avec la vente par correspondance. En revanche, GFK exclut les ventes directes du fabricant au consommateur (par exemple les ventes d'ordinateurs par Dell).

Dans chaque strate est effectuée un tirage aléatoire de points de vente, à probabilités inégales proportionnelles à leur chiffre d'affaires. Le chiffre d'affaires considéré est celui réalisé en biens d'équipement et de loisirs. Avant extrapolation, le chiffre d'affaire des points de vente panélisés représentait il y a deux ans 35% du chiffre d'affaires réel. Il représente aujourd'hui près de 50%. Cette augmentation s'explique par la montée en puissance des grandes chaînes, plus facile à enquêter.

Le nombre de points de vente audités est représenté dans le tableau suivant :

	nombre de points de vente enquêtés	nombre de points de vente total
Hypermarchés	545	1160
Grands magasins	120	135
Grandes chaînes spécialisées	595	884
Traditionnels	215	5008

Le tirage à probabilités inégales a pour conséquence de sur-représenter les gros distributeurs. En contrepartie, peu de petits distributeurs sont sélectionnés, et le poids qui leur est attribué est élevé. Ainsi leur comportement parfois chaotique peut avoir de lourdes conséquences dans les estimations.

Après cette phase de sélection, la totalité des ventes des produits d'équipement et de loisirs est prise en compte dans chacun des points de vente. Le rythme de recueil de l'information varie suivant le produit : de bimestriel pour les secteurs stables (lave-vaisselle) à hebdomadaire pour les secteurs en forte mutation (téléphonie mobile).

Les variables d'intérêt sont les quantités vendues et le prix moyen de chaque modèle pour la période considérée (un bimestre, un mois, une semaine). L'extrapolation de ces variables d'intérêt est différenciée par produit. GFK calcule les coefficients d'extrapolation à l'aide d'enquêtes typologiques régulières, permettant de connaître la part du chiffre d'affaire du point de vente dédié à des ensembles de produits (produits bruns / produits blancs) ou à de grandes familles de produits.

Selon l'importance des cas de non-représentation (régions non couvertes, ventes directes du fabricant, rétrocession à des filiales étrangères...), les taux de couverture après extrapolation oscillent entre 80% et 95%.

1.1.2. Suivi de l'échantillon dans le temps

Le panel de points de vente a été créé en 1979 en collaboration avec la société Sécodip et il est à vocation permanente. Afin de rester représentatif, GFK ajuste au minimum une fois par an les pondérations transversales. Dans la même optique, les nouveaux points de ventes (les naissances) sont régulièrement injectés dans le panel alors que ceux qui font faillite (les morts) en disparaissent. Les cas de non-réponses proviennent essentiellement de problèmes de délais ou de fermeture provisoire de points de vente. Ils sont traités en réattribuant le poids des non-répondants aux répondants de la même strate. Enfin, GFK ne réalise pas de rotation de l'échantillon car le phénomène de lassitude fréquent dans les panels de consommateurs n'existe pas ici.

1.1.3. Enquêtes

A chaque création de produits, les constructeurs transmettent à GFK une documentation technique sur celui-ci. Dès lors GFK réalise centralement la saisie de la nouvelle référence et des caractéristiques techniques dans la base des produits, ce qui génère automatiquement un identifiant appelé *code GFK*. Si une nouvelle référence apparaît dans les ventes d'un point de vente sans qu'elle ait été saisie au préalable (cas des marques de distributeurs), un inspecteur est envoyé dans le point de vente afin de réaliser sa caractérisation par saisie portable. La base des produits, qui est une base nationale, va bientôt laisser place à une base de données européenne, ce qui représentera un gain en qualité et en exhaustivité.

Il existe deux modes de recueil de l'information :

- ?? Cas n°1 : l'enquête comptabilisant les chiffres d'affaires et les ventes chez les détaillants traditionnels
- ?? Cas n°2 : l'enregistrement de sorties de caisses en hypermarchés et grands magasins spécialisés.

Actuellement, le cas n°2 représente 75% de la collecte contre 25% pour le cas n°1. Le cas n°2 est d'ailleurs en constante progression. Il faut noter que dans ce cas, il n'y a pas forcément passage d'enquêteurs dans le point de vente. En effet certains d'entre eux transmettent directement leurs enregistrements par voie informatique.

1.1.4. Disponibilité des données

GFK dispose de données fines au niveau du modèle dans le point de vente. Il ne peut évidemment pas commercialiser ces données pour des questions de confidentialité. Le niveau commercialisable le plus fin correspond aux modèles d'un circuit de distribution pour chacune des 5 régions géographiques GFK.

Les données d'un bimestre sont disponibles 4 à 5 semaines après la clôture du bimestre. Une accélération de ce rythme est envisagée et dépend essentiellement de la rapidité de réponse des points de vente. D'autre part, une mensualisation de l'enquête pour les produits qui nous intéressent (télévisions et lave-vaisselle) semble tout à fait plausible à l'avenir.

1.2. Difficultés de mise en forme des données

1.2.1. Description des fichiers

Les données GFK se présentent sous la forme d'un fichier EXCEL par bimestre pour un produit donné. Ainsi, nous disposons de 30 fichiers pour les lave-vaisselle correspondant aux 30 bimestres d'observation (de février-mars 1996 à décembre 2000-janvier 2001), et de 24 fichiers pour les télévisions (observées de février-mars 1997 à décembre 2000-janvier 2001). Par la suite, nous prendrons la notation « 9603 » pour désigner le bimestre « février-mars 1996 » (le bimestre « décembre 2000 – janvier 2001 » est noté « 0101 »).

Chacun de ces fichiers contient 5 feuilles de calcul. La première représente le total des ventes, et les quatre autres représentent chacun une forme de vente (hypermarchés, grands magasins spécialisés, magasins traditionnels, autres).

Une feuille comptabilise les ventes (du produit considéré) pour un bimestre et une forme de vente donnés. Nous y disposons de la liste des modèles vendus ce bimestre dans l'ensemble des points de vente du panel correspondant à cette forme de vente. A chaque modèle est associé une marque, un code GFK, des Caractéristiques Techniques (CTs), une estimation du nombre d'unités vendues et du prix moyen.

1.2.2. Uniformisation des fichiers Excel en vue d'une importation sous SAS

EXCEL étant peu adapté au traitement de ces fichiers volumineux, nous avons importé les données sous le logiciel SAS. La structure des noms de fichiers et les noms des feuilles pouvaient changer d'une année sur l'autre. L'ordre des variables pouvait également différer. Il a donc fallu procéder à une phase d'uniformisation des fichiers dans le but d'automatiser l'importation.

Une fois l'uniformisation réalisée, les 150 feuilles EXCEL des lave-vaisselle (30 bimestres * 5 feuilles de calcul) et les 120 feuilles des télévisions (24 bimestres * 5 feuilles de calcul) ont été transformées en autant de fichiers SAS.

1.2.3. Création d'un fichier par produit

Les noms de variables étant codés (ex : FEA174), nous leur avons attribué des noms explicites. Puis, nous avons effectué une jointure des 150 (resp. 120) fichiers de lave-vaisselle (resp. télévisions), pour obtenir un fichier de la forme :

Marque	Code GFK	CT1	...	CTn	Forme de vente	Date	Prix moyen	Quantité
--------	----------	-----	-----	-----	----------------	------	------------	----------

Autrement dit, un même modèle vendu sur 6 bimestres différents dans une seule forme de vente apparaît 12 fois dans ce fichier (6 fois pour la forme de vente considérée, et 6 fois pour la forme de vente « totale »). On se retrouve ainsi avec un fichier de 53953 lignes pour les lave-vaisselle et un fichier de 79932 lignes pour les télévisions.

L'information donnée sur le total des ventes n'était pas toujours parfaitement cohérente avec l'information détaillée par forme de vente, en particulier pour les quantités vendues

où l'écart bimestriel pouvait atteindre jusqu'à 11%. Cette information n'a donc pas été utilisée, ce qui réduit le fichier des lave-vaisselle à 34799 lignes et celui des télévisions à 52052 lignes.

Sous cette forme le fichier est peu lisible, nous lui avons préféré un fichier de la forme :

Marque	Code GFK	CT1... CTn	Forme de vente	px9603	qt9603	px9605	qt9605	...	px0101	qt0101
--------	----------	------------	----------------	--------	--------	--------	--------	-----	--------	--------

pour les lave-vaisselle (pour les télévisions, l'historique des prix commence en 9703)

Ainsi le modèle précédent vendu sur 6 bimestres différents dans une forme de vente apparaît 1 seule fois dans le fichier avec 6 prix et 6 quantités renseignés sur 30 (resp. 24) possibles. Cette mise en forme permet de disposer sur une ligne d'une série – un modèle dans une forme de vente – avec sa description et l'historique des prix et quantités par bimestre.

La principale difficulté, ici, est de déterminer ce qu'est un modèle. Lors de notre « mise en séries », nous avons considéré que deux modèles étaient différents lorsqu'une seule des variables suivantes différait (pour une forme de vente donnée) : marque, code GFK, caractéristiques techniques. Ainsi, il y a au maximum 4 séries pour un modèle s'il est vendu dans chaque forme de vente.

A priori, un modèle est entièrement déterminé par la donnée de sa marque et de sa référence. La référence du modèle n'apparaît pas en tant que telle mais est incorporée à la variable *code GFK*, dont elle peut être extraite. Or il arrive fréquemment de trouver une même marque-référence avec des caractérisations différentes : c'est un problème de caractérisations multiples que nous présentons dans la partie suivante.

Un autre problème se pose également : l'apparition de nouvelles variables. Prenons l'exemple des lave-vaisselle. De 1996 à 1998 les modèles vendus sont décrits à l'aide de 15 CTs. En février-mars 1999, 6 nouvelles variables sont introduites. Pour régler ce problème de définition d'un modèle, nous n'utilisons pas toutes les CTs mais seulement celles qui sont communes à toute la période. Nous nous sommes, bien sûr, assurés que la marque, le code GFK et les CTs communes identifiaient de manière unique un modèle et ne généraient pas de caractérisations multiples sur les nouvelles variables.

Les deux fichiers ainsi obtenus contiennent 5310 séries pour les lave vaisselle et 7422 séries pour les télévisions.

1.3. Caractérisation des données

1.3.1. Présentation

L'un des avantages des micro-données est d'offrir une caractérisation détaillée des modèles référencés. Ainsi les téléviseurs se voient associer 21 variables descriptives (y compris la forme de vente et le *code GFK*), et les lave-vaisselle 23¹. Dans la perspective d'une modélisation du prix de ces biens d'équipement en fonction de leurs caractéristiques techniques, il nous a paru essentiel que ces variables soient à même de vérifier des critères de pertinence, d'uniformité, de cohérence et d'exclusivité (une et une seule caractérisation pour une marque-référence donnée).

La mise en défaut de certains de ces critères nous a par conséquent incité à définir une démarche rationnelle de traitement des données.

1.3.2. Détection des problèmes

a) Pertinence

Les biens d'équipement possèdent la particularité d'être extrêmement « techniques », c'est pourquoi il a fallu nous familiariser d'abord avec l'univers de nos deux produits. Si les caractéristiques des lave-vaisselle sont facilement interprétables (hauteur, largeur, nombre de programmes, classes d'efficacité...), il n'en est pas de même pour les téléviseurs. Davantage en proie aux améliorations technologiques, ces derniers conjuguent complexité des options et diversité des appellations constructeurs... à tel point qu'il devient difficile d'appréhender à travers quelques modalités le caractère mouvant de certaines caractéristiques techniques. Les cas les plus épineux concernent le **son** et les **tubes cathodiques**

Nous disposons d'une variable *Dolby*, dont les quatre modalités retracent l'évolution des laboratoires Dolby en matière de reproduction sonore : *Sans Dolby*, *Dolby Surround ou Prologic*, *Dolby Digital*, *Virtual Dolby*. On peut déjà déplorer le fait que les systèmes Surround (1982) et Prologic (1987) ne soient pas différenciés, alors que le second permet d'extraire un canal supplémentaire depuis la source sonore. D'autre part, se restreindre aux marques déposées par les Laboratoires Dolby induit des imprécisions de spécification : ainsi la modalité *Sans Dolby* place des téléviseurs « basiques » au côté de téléviseurs sans logos Dolby, mais équipés de brevets constructeurs (« Incredible Surround » de Sony, « Effet Surround tridimensionnel » de Radiola) qui présentent des propriétés de décodage du son comparables à celles du Dolby Prologic.

En ce qui concerne les tubes cathodiques, la complexité est plus grande encore : « flat square tube », « super flat », de Thomson ; « super trinitron », « FD trinitron », « FD trinitron wega » de Sony ; « blackline S ultra plat », « real flat » de Philips, pour ne citer que les grandes marques... autant de termes pour autant de procédés aux degrés de perfectionnement divers, que l'on considère un constructeur propre ou que l'on compare les constructeurs entre eux. Il est évidemment impossible (autant qu'inexploitable) de créer une variable dont les modalités retracent l'ensemble des techniques existantes. Cependant, la classification qui nous est offerte semble un peu trop simpliste. Elle réduit le nombre de modalités de la variable *Tube* à trois (*Normal* pour les tubes conventionnels, *FST* pour les tubes plats, *RFT* pour les extra-plats), ce qui comme au paragraphe précédent génère des groupes hétérogènes : un tube *FST* Bluesky n'offre vraisemblablement pas la même qualité d'image qu'un tube *FST* Thomson ; le tube *FST* d'un modèle Sony de 1997 est à coup sûr moins perfectionné que le tube *FST* d'un modèle Sony de 2000.

L'exemple de ces deux variables est extrême, et ne saurait concerner l'ensemble des variables. Ce sont d'ailleurs les caractérisations du tube et du son qui pour les téléviseurs sont le plus sujettes aux erreurs. Cependant cette insuffisance du nombre de modalités touche également les variables *Hertz* (absence de la version 100 Hertz numérique, davantage perfectionnée) et *Format* (réunion au sein de la même modalité des options 16/9 et 16/9+, la seconde étant plus évoluée). Mentionnons enfin que certaines variables d'information sont exclues du champ de caractérisation : la puissance musicale, le nombre de hauts parleurs, le nombre de pages télétexte pour les téléviseurs ; l'affichage numérique, l'auto-programmation, l'Aqua Sensor pour les lave-vaisselle. Même si elles sont secondaires, ces options définissent les « petits plus » pouvant influencer sur le prix d'un article.

Tout comme notre Indice de Prix à la Consommation, les micro-données rencontrent donc des difficultés dans leur recherche d'une description fine et pertinente des produits qu'elles audient. Si

¹ Voir le Dictionnaire des Données N°1 [liste des variables GFK] en annexe.

nous avons tenté par la suite d'améliorer la qualité de cette description, nous n'avons évidemment pu y ajouter de nouvelles variables.

b) Uniformité

Tout d'abord, certaines marques présentaient le défaut d'être répertoriées sous deux ou trois orthographes différentes. Bien que mineur (2 marques concernées pour les lave-vaisselle, 9 marques pour les téléviseurs), ce phénomène créait artificiellement plusieurs séries au lieu d'une pour un modèle dans une forme de vente.

Ensuite, deux types de séries sont apparues comme soulevant des problèmes spécifiques de caractérisation. Les premières ne possédaient pas de marque propre mais la modalité *Trade brand*, vocable sous lequel GFK réunit l'ensemble des marques de distributeurs. Les secondes possédaient des chiffres précédés de la mention 'NEU' au sein de la variable *code GFK*, mention repérant une correction d'erreurs réalisée lors du suivi.

Toutes ces séries avaient pour point commun d'être moins bien caractérisées, du fait d'une variable *code GFK* à vide (cas des séries *Trade Brand*) ou incomplète (cas des séries 'NEU'). Cependant elles représentaient une part non négligeable de l'ensemble des données : jusqu'à 6.1% (respectivement 2.4%) du chiffre d'affaires bimestriel² pour les *Trade Brand* téléviseurs (respectivement lave-vaisselle) ; jusqu'à 5.2% (respectivement 3.2%) du chiffre d'affaires bimestriel pour les 'NEU' téléviseurs (respectivement lave-vaisselle).

Enfin, se posait le problème des nouvelles variables introduites en 9903 pour les lave-vaisselle. Deux d'entre elles (*mois_nais* et *an_nais*) remplaçaient une variable déjà existante (*an_intro*). Les autres (*dep_dif*, *cl_power*, *cl_lav*, *cl_sech*), que nous appellerons désormais 'vars_99', étaient vraiment nouvelles, et liées pour la plupart à l'introduction d'une nouvelle norme de codification des lave-vaisselle, obligatoire en France à partir du premier mars 2000. Introduisons maintenant les notations suivante : appelons **pop1** les séries lave-vaisselle « mortes » en 9901 ; **pop2** les séries observées à la fois avant et après 9903 ; **pop3** les séries observées à partir de 9903 seulement. En dépit de notre conservation systématique, lors de la phase de « mise en séries », de la caractérisation la plus récente, trois questions subsistaient :

- ?? Comment choisir entre *an_intro* et *an_nais* pour les séries de la **pop2** en cas de contradiction ?
- ?? Comment faire en sorte que les nouvelles variables ne soient pas à vide pour les séries de la **pop1** ?
- ?? Comment réduire le pourcentage de modalités manquantes pour les classes d'efficacité de la **pop3** (en particulier les modèles récents, apparus en 2000 ou en 2001, et pour lesquels on compte plus de 30% de caractérisations manquantes) ?

c) Cohérence interne et exclusivité

Nous ne pouvions imaginer recenser l'ensemble des informations constructeurs (catalogues, sites Internet) en vue de contrôler la validité des spécifications GFK. Il nous fallait trouver des méthodes systématiques, capables de mettre en évidence les erreurs possibles de codage.

Les premières analyses ont été simplement descriptives. Tout d'abord l'examen des **modalités prises** a révélé plusieurs invraisemblances : des modalités non référencées pour les variables *Nicam* et *Dolby* ; des valeurs aberrantes pour les variables *Kwh* et *Litre* ; un pourcentage élevé de caractérisations manquantes parmi les lave-vaisselle (plus de 6% des séries possédant au moins une modalité manquante, 'vars_99' exclues)... Certaines fréquences agrégées nous ont également semblé bizarres : la variable *Type* par exemple estimait à près de 30% la part des téléviseurs dépourvus de télécommandes.

Un autre mode de contrôle de cohérence a consisté en des **croisements de variables**. Il s'est appuyé sur l'existence de plusieurs caractéristiques techniques à l'intérieur du *code GFK*. Leur extraction a donné lieu à la création, en plus de la *référence*, des variables *TYP* (caractère portable ou table, mono ou stéréo), *DIAG* (diagonale d'écran en cm), *FOR* (format), *TUB* (nature du tube), *REC* (standards de réception) pour les téléviseurs, et de la variable *POS* (caractère posable, intégrable ou encastrable) pour les lave-vaisselle.

Nous avons alors pu comparer les caractérisations issues de « variables jumelles » (*type/TYP*, *pouce/DIAG*, *format/FOR*, *tube/TUB*, *recept/REC*), et voir qu'il n'y avait pas correspondance exacte. En 9811 par exemple, la variable *tube* estime à 69% la proportion de ventes de téléviseurs *FST*, tandis que la variable *TUB* affiche une valeur de 59%.

² Chiffre d'affaires « hors corrections » (ie avant traitement sur les prix).

L'étude progressive des téléviseurs et des lave-vaisselle nous a également appris que pour certains couples de variables, il y avait exclusion (ou correspondance) de modalités : ainsi les téléviseurs dotés de l'option *nicam* sont nécessairement *stéréos* ; une hauteur de 85 cm correspond en règle générale à des lave-vaisselle à poser... Nous avons pu alors établir des tests de détection d'erreurs à un niveau fin (celui de la série).

Notre dernière méthode systématique a reposé sur l'examen des **caractérisations multiples**. Par ce terme on désigne l'ensemble des séries correspondant à une marque-référence pour laquelle la combinaison de modalités prises par les variables descriptives (modalités manquantes y compris) n'est pas unique.

Imaginons que le lave-vaisselle de marque FAURE, de référence 'LVN 165' apparaisse trois fois dans notre fichier : une série pour la forme de vente 'hypermarchés' (codée 54 décibels), une deuxième pour la forme de vente 'magasins traditionnels' (codée 54 décibels toujours), et une troisième pour la forme de vente 'hypermarchés', codée cette fois 52 décibels. On peut comprendre (cf paragraphe 1.2.3.), pourquoi 3 séries ont été retenues dans notre fichier. Ces trois séries font partie des caractérisations multiples du fichier des lave-vaisselle, et doivent être examinées afin de déterminer quel est exactement le niveau sonore du 'LVN 165' de FAURE. Le but est non seulement d'obtenir une caractérisation fiable (et par la suite de mesurer au mieux l'incidence des caractéristiques techniques sur le niveau de prix), mais aussi de pouvoir agréger les deux séries correspondant au réseau des hypermarchés, afin d'obtenir une période de suivi du modèle plus longue (plus proche du cycle de vie effectif du produit), et des prix moyens plus robustes.

Ces caractérisations multiples sont à considérer sérieusement. Toutes variables considérées, elles représentaient 9.8% des séries de téléviseurs (soit 4.6% des marque-référence, et 11.5% du chiffre d'affaires total³). Pour les lave-vaisselle, les proportions passaient à 17.8% des séries, 10.3% des marque-référence et 15.6% du chiffre d'affaires total³ (en excluant du champ des variables les six variables nouvelles introduites en 9903).

1.3.3. Traitement

a) Choix des variables d'intérêt⁴

Toutes les variables GFK n'ont pu être conservées en tant que variables de caractérisation. Notre tâche d'uniformisation et de correction a concerné seulement celles d'entre elles qui caractérisaient techniquement les séries, et qui s'étaient avérées suffisamment fiables. Au rang des variables « non retenues », on distingue donc d'une part les variables de codification interne GFK (*mar*, *numordre*), d'autre part celles qui ne satisfaisaient pas aux critères de cohérence imposés.

Téléviseurs

Après éclatement du *code GFK*, 26 variables étaient à notre disposition. Nous en retiendrons 18 : *marque*, *référence*, *an_intro*, *cartepc*, *combi*, *satellit*, *svhs*, *teletext*, *tuner*, *nicam*, *dolby*, *hertz*, *recept*, *TUB*, *FOR*, *TYP*, *DIAG*, *fv*.

Dans la plupart des cas, les variables issues du *code GFK* (celles notées en majuscules) ont été préférées à leur équivalent de la nomenclature (*tube*, *format*, *type*, *pouce*). La raison en est qu'elles sont davantage cohérentes, qu'il s'agisse de calculs de fréquences globaux, de caractérisations multiples (elles en occasionnent moins), ou de vérifications catalogues occasionnées par une non-correspondance entre « variables jumelles » (voir paragraphe 1.3.2. c)). La seule exception concerne la variable *REC*, moins riche que *recept* en modalités offertes.

Lave-vaisselle

Nous avons retenu finalement 23 variables descriptives, sur les 24 disponibles après éclatement du *code GFK* : *marque*, *référence*, *POS*, *hauteur*, *largeur*, *couvert*, *nbprog*, *pan_reg*, ***k1***, ***I1***, ***k2***, ***I2***, *decibel*, *nb_tempe*, *type*, *dep_dif*, *cl_power*, *cl_lav*, *cl_sech*, ***an_naiss***, *fv*, ***code***.

Nous avons été amenés à remplacer la variable *kwh* par *k1* et *k2*, la variable *litre* par *I1* et *I2*, les indices « 1 » correspondant aux consommations pour l'ancienne norme (ie avant la réglementation du premier mars 2000), et les indices « 2 » aux consommations pour la nouvelle norme. En effet d'une norme à l'autre le programme de référence peut changer, par conséquent la valeur des consommations aussi. Notre hypothèse fut de poser que les séries pour lesquelles les classes

³ Chiffre d'affaires « hors corrections » (ie avant traitement sur les prix).

⁴ Voir le Dictionnaire des Données N°2 [liste des variables retenues] en annexe.

d'efficacité n'étaient pas renseignées étaient codifiées selon l'ancienne norme, et les autres (minoritaires : 890 seulement sur les 5310) selon la nouvelle.

La variable *an_naiss* représente l'année d'introduction du modèle, et a été construite à partir des variables *an_intro* et *an_nais* (dès l'étape de la « mise en séries » en vérité). Elle est égale à *an_intro* pour les séries de la **pop1** (voir paragraphe 1.3.2. b)), à *an_nais* pour les séries de la **pop3**, et au minimum (*an_intro*, *an_nais*) pour les séries de la **pop2**. Le choix de cette formule permet de régler les cas de non égalité entre les deux variables. Il est légitimé par le fait que dans tous ces cas, l'année d'introduction la plus récente était postérieure au premier bimestre d'observation de la série (chose évidemment impossible).

La présence dans nos fichiers de plusieurs occurrences d'un même lave-vaisselle, décliné en plusieurs coloris, nous a incité à créer de surcroît la variable *code*. Cette variable correspond en fait à la *référence* du produit, privée des lettres qui pour sa marque représentent la couleur (*BL* ou *WH* pour *blanc* chez la plupart des constructeurs, *X* pour inox chez Vedette...). Cette variable *code* s'est en fait avérée très utile, tant pour l'examen des caractérisations multiples que pour l'imputation de modalités incorrectes ou manquantes.

Remarque : Les variables « non retenues » n'ont pas été éliminées du fichier. En effet, elles nous ont servi en cas de mise en défaut des variables retenues, comme nous allons le voir maintenant.

b) Uniformisation des *Trade Brand* et des *NEU*

Même au sein des produits dits «MDD», il existe une hiérarchie de gammes. C'est pourquoi il est intéressant de pouvoir différencier les *Trade Brand* les uns des autres. La différenciation par les marques put s'effectuer pour la totalité des 97 séries de téléviseurs, et pour 42 des 44 séries de lave-vaisselle, ce grâce à la variable *numordre*. En effet, les numéros d'ordre attribués à ces séries *Trade Brand* correspondaient à une marque unique pour les séries «classiques» dont la marque était renseignée. On a donc pu facilement imputer des noms de marque tels que *BLUESKY*, *FIRSTLINE* ou encore *NOGAMATIC* aux séries concernées. Pour les variables de caractérisation, la démarche fut moins aisée. Toutes les variables issues du *code GFK* étant à vide, nous avons choisi de leur attribuer des valeurs en s'appuyant sur les modalités prises par leurs variables jumelles. Ainsi les valeurs de *TYP* furent par exemple obtenues à partir de la variable *type*, en associant à chaque modalité de *type* l'occurrence la plus fréquente de *TYP* sur l'ensemble des observations. En ce qui concerne les lave-vaisselle, nous ne disposions pas de « variable jumelle ». C'est alors la hauteur qui a servi à l'imputation de *POS* (selon la règle : *85 cm* -> *FS*, *82 cm* -> *IE*). Pour la référence enfin aucun élément ne permettait la reconstitution, elle est donc restée absente des enregistrements.

Les modèles dits '*NEU*' représentent pour les téléviseurs comme pour les lave-vaisselle environ 2% des séries, soit 4% des marque-référence. Ils constituent un terrain propice à un **manque d'information** (toutes les variables extraites du *code GFK* sont manquantes, excepté la *référence*) et à la **caractérisation multiple** : la plupart des marque-référence concernées apparaissent par ailleurs sous une forme '*non NEU*' (dite « complète ») dans les tables.

Pour les séries possédant des « équivalentes » (ie des séries de même marque-référence) « complètes », il fut simple de remédier aux défauts mentionnés : nous leur avons attribué la caractérisation de ces équivalentes. Ceci concernait 64% des séries de téléviseurs, et 79% des séries de lave-vaisselle. Pour ces derniers, nous avons pu encore améliorer les choses en remplaçant le critère d'équivalence 'marque-référence' par le critère 'marque-code' (la caractérisation technique ne dépendant pas de la couleur du modèle). Le pourcentage de séries '*NEU*' pour lesquelles on peut trouver une caractérisation complète est alors passé à 93%.

Pour les séries ne possédant pas « d'équivalentes complètes », nous avons choisi d'avoir recours ou bien à une recherche catalogue (puisque la référence est connue), ou bien à un processus d'imputation semblable à celui des *Trade Brand*.

c) Uniformisation des variables lave-vaisselle

Au paragraphe 1.3.2. b) figurent trois questions soulevées par l'apparition de nouvelles variables descriptives en milieu de période d'étude. La réponse à la première d'entre elles a été la création de la variable *an_naiss*, comme nous l'avons vu plus haut.

Au lieu de remplacer systématiquement les valeurs à vide des '*vars_99*' par des modalités manquantes pour les séries de la **pop1**, nous leur avons attribué lorsque c'était possible les modalités de leurs « équivalentes » des **pop2** et **pop3**, en s'appuyant sur le critère d'équivalence 'marque-code'. Ainsi nous avons pu enrichir la caractérisation de nos données (principalement pour la variable *dep_dif*) et harmoniser les codes de non-spécification entre populations.

Il était naturel que les variables *cl_power*, *cl_lav*, *cl_sech* présentent un taux de remplissage si faible (seules 17% des séries étaient renseignées). La nouvelle norme européenne prenant effet au premier mars 2000, les constructeurs n'ont en général défini des classes d'efficacité qu'à partir de 2000, sans revenir sur la caractérisation de leurs anciens modèles. Ainsi, on comprend que les séries dont l'année d'introduction est antérieure à 2000 soient très rarement caractérisées. En revanche, nous nous sommes attachés à l'amélioration de la caractérisation des modèles récents : c'est important lorsqu'on sait qu'au sein du modèle économétrique utilisé pour le traitement des remplacements dans l'IPC, le pouvoir explicatif⁵ des classes d'efficacité énergétique et de séchage atteint 33%. Cette amélioration a reposé sur l'examen des catalogues et des sites des constructeurs les plus connus. Elle s'est accompagnée du remplissage des variables *k2* et *l2* (aux modalités différentes parfois de celles de *k1* et *l1*). La plus-value qu'apporte cette recherche est intégrée aux calculs d'évaluation globaux, exposés en section 1.3.5.

d) Erreurs de codage : corrections

Nous nous bornerons ici à présenter les divers procédés qui nous ont servi à corriger les erreurs de codage que révélèrent l'examen des tests de cohérence entre variables croisées, des modalités manquantes ou aberrantes, et des caractérisations multiples. D'autres erreurs furent détectées de façon non systématique, au gré de l'analyse d'une marque ou d'une variable de caractérisation précise. Mais il est certain que malgré cela nous n'avons pu les repérer toutes.

Nous pouvons donc mentionner au rang des méthodes de correction :

- ?? Le recours à la caractérisation des modèles dans la base de données IPC (mais bien souvent les modèles à problèmes sont des modèles peu courants, et n'appartiennent pas au champ de l'indice).
- ?? L'examen des catalogues constructeurs, ou la visite de leurs sites. L'inconvénient de cette méthode est qu'elle est fastidieuse, applicable uniquement aux grandes marques et à leurs modèles les plus récents. L'avantage est qu'elle permet de contrôler la totalité des modalités associées à une marque-référence.
- ?? Le recours à la structure des références. Par exemple les téléviseurs '28DK24E' et '21DK24E' de Thomson ne diffèrent que par leur diagonale d'écran. Si l'on hésite entre les formats 4/3 et 16/9 pour le premier modèle (du fait d'une double caractérisation), et que le second est décrit de façon unique comme étant un téléviseur 4/3, on imputera la valeur 4/3 à la variable *FOR* des séries '28DK24E'.
- ?? La présence d'extensions spécifiques au sein de la référence. Pour certaines marques de téléviseurs, l'existence de termes tels que 'NX', 'NTX', 'NT' sont révélateurs de la possession des options *nicam* et *télétexte*.
- ?? Les variables non retenues peuvent également servir comme aide au choix : tel modèle dont l'une des séries est dite *FST*, et d'autres *RFT*, sera considéré *FST* si c'est la modalité qu'affiche la variable *tube* pour l'ensemble des séries.
- ?? Un autre de nos critères de choix a été de privilégier, en cas de caractérisation multiple, les séries observées le plus longtemps (ie les séries les plus « stables ») en faisant l'hypothèse qu'elles possèdent une meilleure caractérisation que celles ayant fait l'objet d'un ou deux relevés seulement.
- ?? De même lorsqu'une modalité l'emportait largement sur les autres, parmi les caractérisations multiples d'une même marque-référence, nous avons privilégié cette dernière.
- ?? Enfin le cas particulier des années d'introduction est à mentionner. En cas de caractérisation multiple pour une marque-référence, nous avons procédé comme pour la création de la variable *an_naiss* des lave-vaisselle, ie que nous avons systématiquement choisi la date la plus ancienne.

Remarque : La correction des séries dites « caractérisations multiples » nous a permis d'obtenir l'exclusivité des spécifications. A l'exception de quelques cas extrêmes, désormais une marque-référence est caractérisée de façon unique dans nos fichiers. Pour les lave-vaisselle, nous avons étendu cette exigence au critère d'équivalence 'marque-code', sauf dans deux cas précis : le premier lorsqu'il nous était impossible de déterminer quelle était la « bonne » référence parmi celles possédant le même code ; la seconde pour la variable *an_naiss*, qui pour un code donné présentait une extrême variabilité d'une référence à l'autre (voir fin du paragraphe 1.3.5.). Dans ces deux cas, nous nous sommes donc satisfaits d'une cohérence au sein de la marque-référence.

⁵ Ce pouvoir explicatif est mesuré par la différence de R^2 entre modèle complet et modèle restreint.

Cette extension du critère d'équivalence aura grandement aidé à la détection d'erreurs de codage : ainsi pour les variables de consommation des lave-vaisselle, nous avons détecté 51 marque-référence à problème, contre 24 à partir du seul examen de la marque-référence.

Ajoutons enfin que nous aurions pu étendre le critère d'équivalence non plus à la référence ou au code, mais à la « famille » du modèle (par exemple les 'VIP1', 'VIP2', 'VIP3' de Sauter). Nous avons contrôlé l'unicité de caractérisation au sein de ces familles lorsque le cas s'est présenté à nous, mais il aurait été impossible de systématiser cette méthode sans tomber dans un examen exhaustif des séries.

1.3.4. Agrégation

L'objectif des phases de détection d'erreurs, puis de traitement, était de respecter au mieux les critères de cohérence énoncés au paragraphe 1.3.2. . En particulier, la recherche d'exclusivité dans la caractérisation a généré des séries identiques en tous points pour les variables d'intérêt. Il convenait d'agrèger ces séries, qui en réalité n'avaient été séparées qu'artificiellement (exception faite des *Trade Brand*), afin de représenter ce qu'était leur véritable cycle de vie. Nous avons donc défini une procédure d'agrégation, regroupant les séries *non-Trade Brand*⁶ identiques (même marque, même référence, même forme de vente, mêmes modalités descriptives), et calculant pour chaque bimestre des prix moyens pondérés.

Nous avons ainsi réduit le nombre de séries des téléviseurs à 7123, et celui des lave-vaisselle à 4945. Ceci s'est accompagné d'une augmentation de la durée d'observation moyenne des séries.

1.3.5. Evaluation

Il est important de donner une idée des efforts et des gains occasionnés par le repérage des problèmes, leur traitement et la procédure d'agrégation. Pour cela nous avons retenu trois outils d'évaluation.

?? Evaluation du travail de correction

Le tableau ci-dessous fournit quelques chiffres concernant les marque-référence dont au moins l'une des modalités de variable d'intérêt a fait l'objet de correction, y compris imputation suite à une valeur manquante (exception faite des variables *marque*, *référence*, *code*, *fv*). Le travail d'uniformisation des 'vars_99' n'est pas pris en compte ici.

	Marque-référence corrigées	Min de modalités corrigées par marque-référence	Max de modalités corrigées par marque-référence	Moyenne de modalités corrigées par marque-référence
Téléviseurs	20.2%	1	5	1.09
Lave-vaisselle	29.0%	1	11	1.85

Il apparaît que les lave-vaisselle ont nécessité davantage d'imputations que les téléviseurs : plus de marque-référence concernées ; plus de modalités corrigées en moyenne pour chacune de ces marque-référence. Ceci est certainement lié au surcroît de corrections induit par les modalités manquantes (problème inexistant pour les téléviseurs). Le recours à la variable *code*, en favorisant la détection d'erreurs, peut constituer un facteur supplémentaire d'explication. Il est à noter que certaines variables « concentrent » les défauts de caractérisation. C'est le cas des variables correspondant au son et au tube cathodique pour les télévisions, aux dimensions et aux consommations d'eau et d'électricité pour les lave-vaisselle.

?? Variables manquantes : cas des lave-vaisselle

Le tableau suivant recense les séries et marque-référence possédant au moins une modalité manquante pour les variables d'intérêt, exception faite de la *référence* et du *code* (afin de ne pas prendre en compte les séries *Trade Brand*). Les 'vars_99' font l'objet d'un traitement particulier : seules sont comptabilisées les modalités manquantes des séries pour lesquelles *an_naiss* est postérieur à 2000.

⁶ Pour les séries *Trade Brand*, une procédure analogue d'agrégation avait été effectuée dès l'étape de la mise en forme des fichiers, en utilisant comme critère de regroupement l'égalité des marques et des variables de caractérisation (y compris les variables non retenues). La référence ne put y être incluse, puisque pour ces séries elle est absente.

	Chiffre d'affaires ⁷	Marque-référence	Séries	Nombre moyen de modalités manquantes
Avant correction	6.6%	11.3%	9.1%	3.05
Après correction	2.9%	7.0%	5.2%	3.47

Les modèles qui conservent des modalités manquantes sont majoritairement des modèles dont les marques sont de gammes inférieures, et dont la spécification est absente pour un grand nombre de variables. C'est la raison pour laquelle le nombre moyen de modalités manquantes augmente après correction, et que la part en chiffre d'affaires de ces modèles reste faible, malgré des pourcentages de non-caractérisation relativement élevés.

?? Caractérisations multiples

Les chiffres présentés maintenant s'intéressent à l'évaluation des caractérisations multiples sur variables d'intérêt, avant et après correction. On a pris soin d'exclure la population des *Trade Brand* de l'étude⁸. Le critère d'équivalence retenu est la marque-référence.

Pour les lave-vaisselle, les chiffres entre parenthèses correspondent aux pourcentages de caractérisations multiples hors 'vars_99'.

	Séries téléviseurs	Marque-référence téléviseurs	Chiffre d'affaires téléviseurs ⁶	Séries lave-vaisselle	Marque-référence lave-vaisselle	Chiffre d'affaires lave-vaisselle ⁶
Avant correction	7.1%	3.5%	8.3%	34.5% (16.9%)	24.3% (10.0%)	34.9% (14.4%)
Après correction	0.3%	0.3%	0.1%	0.0%	0.0%	0.0%

Le taux d'erreurs associées à des caractérisations multiples a donc été significativement réduit. Les quelques cas subsistant au sein des téléviseurs concernent des modèles de marques peu connues, pour lesquelles nous ne disposons pas d'information catalogue, et qui sont absentes du champ de l'IPC. C'est pourquoi nous avons décidé alors de conserver l'ensemble des caractérisations proposées.

Si les lave-vaisselle satisfont complètement à l'exigence d'une caractérisation unique pour une marque-référence donnée, certaines incohérences existent encore si l'on passe au critère d'équivalence marque-code. Comme nous l'avons vu précédemment, ceci concerne surtout la variable *an_naiss*, pour laquelle 82 marque-code possèdent plus d'une modalité. Si l'on peut comprendre que certaines versions apparaissent après d'autres (par exemple les modèles « graphites » ou « alus »), des décalages de 5 ans paraissent suspects. On peut alors penser que pour GFK, l'année d'introduction d'un modèle correspond en fait à l'année de sa première observation.

Enfin si nos données obéissent désormais à davantage de cohérence, nous avons conservé des incertitudes sur certaines séries, sans pouvoir les modifier faute d'information extérieure. Pour les téléviseurs 14 marque-référence sont concernées (sur 3252), et pour les lave-vaisselle 48 (sur 2053).

1.4. Traitement des incohérences sur les prix

Afin de simplifier la notation dans cette section, on remplacera pour une série le terme « prix moyen » par le terme « prix ».

Un rapide examen des prix montre plusieurs incohérences : un prix de 20F pour une série un certain bimestre ; des prix multipliés ou divisés par 60 d'un bimestre à l'autre. Tout comme pour les caractéristiques techniques, il convient de corriger ces incohérences. Cependant, la multitude des prix présents (51883 pour les télévisions et 34723 pour les lave-vaisselle) ne permettant pas un traitement manuel, elle oblige à créer un algorithme de repérage et de traitement des prix aberrants.

1.4.1. Repérage et traitement des prix aberrants

En se plaçant au niveau de la série, on définit deux types d'écarts relatifs de prix, notés eA et eB , entre deux observations consécutives. Si l'on note 0 et 1 les deux dates d'observations consécutives, on a :

⁷ Chiffre d'affaires « hors correction » (ie avant traitement sur les prix).

⁸ Ces deux spécifications justifient les légères différences observées avec les chiffres du paragraphe 1.3.2. c).

$$eA_i = \left| \frac{\text{prix}_i - \text{prix}_0}{\text{prix}_0} \right| \quad \text{et} \quad eB_i = \left| \frac{\text{prix}_i - \text{prix}_1}{\text{prix}_1} \right|$$

Exemple : Prenons une série (un modèle dans une forme de vente) vendue entre 9705 et 9809 avec visiblement un prix aberrant en 9711.

	9703	9705	9707	9709	9711	9801	9803	9805	9807	9809
Prix (en F)	-	5 000	5 100	4 990	2 400	4 900	4 900	-	4 800	4 790
eA (en %)	-	-	2,00%	2,16%	51,90%	104,17%	0,00%	-	2,04%	0,21%
eB (en %)	-	-	1,96%	2,20%	107,92%	51,02%	0,00%	-	2,08%	0,21%

En 9805, la série n'est pas observée, c'est un phénomène de discontinuité⁹. Dans ce cas, les écarts en 9807 sont calculés à partir de la dernière observation, soit celle de 9803. Si la durée de discontinuité est longue cela conduit à comparer des observations très éloignées dans le temps mais cette solution (choisir de comparer plutôt que d'ignorer) paraît préférable.

Une fois ces calculs effectués, on définit des seuils pour eA (noté *seuilA*) et pour eB (noté *seuilB*), à partir desquels on juge qu'il y a existence de prix aberrant.

Dans l'exemple, si on se fixe pour *seuil* (A et B) 100%, eA et eB mettent en évidence un prix aberrant. Si on choisit 105%, seul eB le désigne. Et si on prend 110%, cette série ne présente pas de prix aberrant. Ceci montre l'importance du choix des seuils A et B. Nous reviendrons sur ce choix ultérieurement.

Pourquoi avoir choisi deux types d'écarts relatifs ? En réalité, ils représentent deux types d'évolutions différentes. En effet, eA est plutôt sensible aux fortes hausses de prix et eB aux fortes baisses de prix. Si l'on suppose que 9711 est le dernier bimestre d'observation de la série, et qu'on prend *seuilA=seuilB=100%*, on voit que seul eB est capable de montrer l'existence d'un prix aberrant. Inversement, si la série commençait en 9711, seul eA nous aurait alerté.

La procédure de correction est séquentielle : elle traite successivement les prix aberrants mis en lumière par eB puis par eA. Détaillons la méthode pour eA, le traitement sera analogue pour eB.

Supposons que pour une série, on trouve une date *i* telle que $eA_i > \text{seuilA}$. Si *j* désigne la dernière date d'observation avant la date *i*, on a :

$$eA_i = \left| \frac{\text{prix}_i - \text{prix}_j}{\text{prix}_j} \right| > \text{seuilA}$$

Cela signifie que soit le prix_i soit le prix_j est aberrant. Pour déterminer le « mauvais » prix, on calcule la moyenne pondérée partielle des prix de la série privés de prix_i et de prix_j . Autrement dit,

$$\text{moyennepartielle} = \frac{\sum_{\substack{k=1 \\ k \neq i, j}}^N \text{quantité}_k \cdot \text{prix}_k}{\sum_{\substack{k=1 \\ k \neq i, j}}^N \text{quantité}_k}$$

où N est le nombre total de bimestres (30 pour les lave-vaisselle et 24 pour les télévisions), et où l'on a par convention $\text{quantité}_k=0$ et $\text{prix}_k=0$ si aucune observation n'existe à la date k.

Il est plus intéressant de pondérer les prix par les quantités car plus un modèle est vendu plus son prix est stable (propriété de la moyenne). Ainsi une moyenne pondérée semble préférable à une moyenne simple.

Le « mauvais » prix est alors le prix le plus éloigné de cette moyenne partielle, ce que l'on peut formaliser de la manière suivante :

⁹ Le traitement des discontinuités ne sera pas abordé ici. Nous adapterons en conséquence nos formules d'indices de prix, tout en ayant conscience que le choix s'en trouve restreint.

si $|prix_i - moyenne partielle| \geq |prix_j - moyenne partielle|$ alors $price_i$ est le mauvais prix.
 si $|prix_i - moyenne partielle| \geq |prix_j - moyenne partielle|$ alors $price_i$ est le mauvais prix.

Un problème évident se pose lorsque le modèle n'est vendu que sur deux bimestres et qu'un des écarts relatifs (eA ou eB) est supérieur à son seuil. Dans ce cas, on ne peut pas juger quel est le mauvais prix, on supprime donc cette série.

Un problème similaire est celui des séries observés un seul bimestre : on ne peut pas juger si leur prix est aberrant ou non. On décide, sauf exception, de conserver la ligne. De toute façon, ces séries ne seront pas utilisées dans des calculs d'indices mais éventuellement dans des modèles de régressions.

Le mauvais prix étant maintenant détecté, on modifie celui-ci en utilisant l'observation précédente et l'observation suivante :

$$prix_{modifié} = \frac{quantité_{précédent} \cdot prix_{précédent} + quantité_{suivant} \cdot prix_{suivant}}{quantité_{précédent} + quantité_{suivant}}$$

Cette imputation se réduit au $price_{suivant}$ s'il n'y a pas de $price_{précédent}$ et réciproquement (il en existe au moins un sur les deux !).

Reprenons l'exemple ci-dessus, et réalisons le traitement avec un $seuilA$ (pour l'instant arbitraire) de 100%.

On a : $eA_{9801} = 100\%$.

On retire donc $price_{9711}$ et $price_{9801}$ pour calculer la moyenne partielle des prix restants, on trouve 4 930F.

Pour l'exemple, nous avons supposé les quantités constantes sans perdre de généralité.

$|price_{9711} - moyenne\ partielle| = 2530$ et $|price_{9711} - moyenne\ partielle| = 30$

Le prix aberrant se trouve donc en 9711.

Sa nouvelle valeur devient : $price_{9711} = \frac{price_{9709} + price_{9801}}{2} = 4945$ F.

Le traitement décrit est répété au niveau de la série pour toute date i telle que $eA_i > seuilA$.

Une fois ces traitements successifs réalisés, on compare le nombre de prix corrigés au nombre de prix observés pour chaque série. Si on a corrigé plus de la moitié des prix de la série, on la supprime car elle présente trop d'incohérences.

Autrement dit, si le ratio $\frac{\text{nombre de prix corrigés}}{\text{nombre de prix observés}}$ est supérieur à 0,5 alors la série est supprimée.

1.4.2. Le choix des seuils

Une question délicate en suspens reste le choix de $seuilA$ et $seuilB$. Si ces seuils sont trop élevés, certains prix aberrants ne seront pas traités (situation permissive). Si ces seuils sont trop bas, on prend le risque de corriger des prix justes (situation stricte). Pour les choisir au mieux, nous avons utilisé la définition des valeurs aberrantes dans un box plot¹⁰, selon le processus suivant :

Pour chaque série, on détermine le maximum de eA et celui de eB , notés $eAmax$ et $eBmax$:

$$eAmax = \max_{i \in \{1..N\}} eA_i \quad \text{et} \quad eBmax = \max_{i \in \{1..N\}} eB_i$$

Puis on calcule, pour chacune de ces variables, les premier et troisième quartiles $Q1$ et $Q3$. Ceci permet de définir un seuil à partir duquel les valeurs sont jugées aberrantes par rapport à la distribution.

$$SEUIL = Q3 + 3 * (Q3 - Q1)$$

Les valeurs des seuils sont résumées dans le tableau suivant :

¹⁰ voir P. TASSI, *Méthodes statistiques*. Economica, deuxième édition 1989

	télévisions	lave-vaisselle
<i>seuilA</i>	107,675 %	57,217 %
<i>seuilB</i>	117,093 %	60,000 %

On peut se demander s'il est légitime de différencier les valeurs de *seuilA* et *seuilB* au sein d'un même produit, ou encore d'avoir des seuils différents d'un produit à un autre.

SeuilA et *seuilB* peuvent être différents car les variables *eA* et *eB* ne réagissent pas de la même manière aux fortes évolutions de prix : *eA* est sensible aux fortes hausses et *eB* sensible aux fortes baisses. Le prix des produits considérés ayant une tendance à la baisse, il paraît donc plus aberrant d'assister à une forte hausse qu'à une forte baisse. Ceci explique que le *seuilA* soit dans les deux cas moins élevé que le *seuilB*.

Les seuils des télévisions sont plus élevés que les seuils des lave-vaisselle car les prix observés sont beaucoup plus volatils pour ce premier produit. En fait, une évolution de prix de 60% est cohérente pour les télévisions alors qu'elle paraît totalement excessive sur le marché plus stable des lave-vaisselle.

Il serait intéressant d'étudier l'impact du choix des seuils sur les calculs d'indices ou sur les modélisations hédoniques. En particulier, on pourrait étudier la stabilité des statistiques selon le caractère strict ou permissif de ces seuils.

1.4.3. Impact des traitements réalisés

La table initiale – celle issue de l'apurement des CTs – comporte 4 945 séries et 34 723 prix pour les lave-vaisselle. La table des télévisions se compose de 7 123 séries et de 51 883 prix.

Le nombre de séries présentant au moins un prix aberrant (avec les seuils fixés) est représenté dans le tableau suivant. Il distingue les séries où à la fois *eA* et *eB* dépassent leur seuil de celles où seul l'un des deux écarts dépasse son seuil.

	télévisions	lave-vaisselle
<i>eA</i> et <i>eB</i>	57	47
<i>eA</i> seul	62	47
<i>eB</i> seul	110	90
Total	229	184

Nous avons vu que les traitements sur les prix peuvent donner lieu à la suppression de séries, ce pour deux raisons principales : la première, lorsqu'une série ne possède que deux observations dont l'une est aberrante, la seconde lorsqu'on corrige plus de la moitié des prix de la série.

Le tableau suivant comptabilise les « pertes » engendrées par la correction des prix :

		table initiale	table finale	<i>perte</i>
télévisions	nombre de séries	7 123	7095	28
	nombre de prix	51 883	51820	63
lave-vaisselle	nombre de séries	4 945	4 931	14
	nombre de prix	34 723	34 663	60

On impute donc finalement 201 séries pour les télévisions (229-28) et 170 séries pour les lave-vaisselle (184-14). La distribution de ces séries selon le nombre de prix corrigés est celle-ci :

	télévisions	lave-vaisselle
1	177	151
2	22	16
3	1	2
4	0	1
5	1	0
Total	201	170

Ainsi, un rapide comptage montre que l'on corrige 229 prix pour les télévisions et 193 prix pour les lave-vaisselle. Cela peut paraître relativement peu, mais le choix de seuils de correction plus stricts aurait conduit à un nombre de corrections beaucoup plus élevé.

La correction des prix aberrants est primordiale pour le calcul d'indices de prix. Un seul prix aberrant, en particulier s'il se situe à la période de base, peut fausser totalement leur valeur.

Une autre approche, plus systématique, consiste à supprimer la série entière dès qu'elle présente un prix aberrant. Cette approche est dommageable car on perd de la « bonne » information. Dans l'exemple précédent, pour un prix aberrant, on aurait supprimé 7 « bons » prix et toutes les caractéristiques techniques de la série. Si on avait supprimé ces séries au lieu de les imputer, on aurait perdu **4,1%** du chiffre d'affaires¹¹ des lave-vaisselle, et **1,5%** de celui des télévisions.

On aurait alors pu simplement supprimer la mauvaise valeur, mais cela génère des discontinuités fictives, également gênantes pour le calcul d'indices. La technique d'imputation par la moyenne des observations adjacentes nous a paru convenable même si elle présente l'inconvénient de tenir compte du futur. Dans un objectif de production, cette technique devrait donc être adaptée.

¹¹ Il s'agit du chiffre d'affaire réalisé sur toute la période d'observation.

2. LES DONNEES IPC

2.1. Présentation des Biens Durables dans l'indice des prix à la consommation

2.1.1. Un secteur spécifique

La construction de notre Indice de Prix repose sur l'observation de produits, choisis dans les points de vente de l'échantillon par les enquêteurs. Chaque mois sont relevés le prix et les caractéristiques techniques de ces produits, que nous appellerons désormais séries élémentaires. Cette notion IPC de série élémentaire (un représentant dans un point de vente d'une marque-référence) est à distinguer de la notion GFK de série (une marque-référence dans une forme de vente).

Pour la plupart des variétés, les enquêteurs accomplissent eux-mêmes les changements de séries élémentaires en cas de disparition en cours d'année. Pour cela ils s'appuient sur l'examen des feuilles de collecte répertoriant en plus du prix les caractéristiques techniques des produits enquêtés. Les variétés Biens Durables quant à elles ont la particularité de posséder un système de gestion des remplacements centralisé : sur le terrain, les enquêteurs qui doivent opérer un remplacement proposent deux produits de substitution ; puis l'équipe « Biens Durables » de la direction générale de l'INSEE choisit le remplaçant parmi ces deux propositions. Une telle méthode permet le contrôle des remplacements, grâce à une équipe spécialisée disposant de données constructeurs (catalogues, sites Internet, contacts avec les services commerciaux) s'ajoutant aux données de terrain. Elle donne également lieu à un archivage complet des modèles suivis dans l'IPC, et de leurs caractéristiques techniques. Cet archivage est effectué par l'équipe « Biens Durables » qui vérifie et complète la description des modèles retenus dans l'échantillon. Cette méthode, qui assure la qualité et l'homogénéité de la description des produits, facilite grandement l'exploitation des données sur séries élémentaires contenues dans le système informatique national.

L'Indice des Prix à la consommation distingue les variétés homogènes (constituées de produits considérés équivalents) des variétés hétérogènes (constituées de produits dont les prix et évolutions de prix sont difficilement comparables). Les formules d'indices retenues dans les deux cas ne sont pas les mêmes. Nous allons présenter maintenant celles qui sont associées aux variétés hétérogènes, les variétés du secteur Biens Durables (en particulier les téléviseurs et les lave-vaisselle) appartenant à ce dernier groupe.

2.1.2. Formule de calcul d'un indice de variété hétérogène

Les séries élémentaires d'une variété sont réparties géographiquement sur le territoire dans $a=96$ agglomérations.

Soit j une de ces agglomérations et N_j le nombre de séries élémentaires de l'agglomération j .

Soit s_j une série élémentaire de l'agglomération j .

On désigne par 0 la période de base et par 1 la période courante.

On définit l'indice entre 0 et 1 d'une variété-agglomération (var-agglo) comme suit :

$$I_j = \sqrt[N_j]{\prod_{s_j=1}^{N_j} I_{s_j}} \quad \text{où } I_{s_j} = \frac{\text{prix}_{s_j}^1}{\text{prix}_{s_j}^0}$$

C'est une moyenne géométrique équipondérée des indices des séries élémentaires. Il faut noter que cette équipondération est une pondération en valeur.

L'indice de variété s'obtient par agrégation des indices de var-agglos.

$$I = \prod_{j=1}^a \text{pond}_j^0 * I_j$$

où pond_j^0 représente la part en valeur de consommation de l'agglomération j ($\prod_{j=1}^a \text{pond}_j^0 = 1$).

Pour une variété donnée, toutes les agglomérations ne sont pas forcément enquêtées. Dans ce cas précis, par convention $I_j=0$.

On voit que ni la forme de vente, ni la notion de modèle ne rentrent dans la formule de calcul de l'indice, ce qui représente une difficulté pour l'agrégation de ces données au format série (au sens vu précédemment).

2.2. Traitement des données IPC

2.2.1. Rapatriement des données

Nous avons vu précédemment que la « centralisation » spécifique aux Biens Durables permettait une exploitation plus aisée des données du système informatique national. Il a fallu néanmoins passer par la jointure sous SAS de plusieurs tables ORACLE pour réunir les informations relatives à l'agglomération d'attache, la caractérisation technique, les divers prix de chaque série élémentaire. Ainsi nous avons obtenu deux tables (une pour chaque produit), de formes comparables à celles des tables GFK. Pour les lave-vaisselle par exemple, cela donne :

Marque	Référence	CTs	Fv	Id agglo	Px base	px9602	...	px0101	pond96	...	pond01
--------	-----------	-----	----	----------	---------	--------	-----	--------	--------	-----	--------

La différence avec les tables GFK est qu'ici une ligne correspond à une série élémentaire, non plus à une série. De plus, se substituent aux critères de quantité les pondérations des 96 agglomérations, pour les différentes années d'étude.

2.2.2. Jointure entre l'IPC et GFK

Afin d'établir des comparaisons fines entre nos deux sources de données, il était important de relier les modèles de l'IPC à ceux de la base GFK (supposée exhaustive ou presque).

Au niveau des marques, le problème est assez simple : toutes les marques présentes dans l'IPC (28 pour les télévisions et 24 pour les lave-vaisselles) se retrouvent dans la base GFK. Seule la marque SINGER, suivie dans l'IPC et non référencée dans GFK, fait exception.

Au niveau des modèles, le problème est plus délicat car toutes les références ne sont pas directement comparables entre elles. En effet, seul 2/3 des références de l'IPC se retrouvent exactement dans les tables GFK. Les autres ne trouvent pas de correspondance automatique du fait de problèmes d'orthographe (exemple : 51-293T et 51293T), d'extensions (52TA1262 et 52TA1262A), d'ordre (FAVORIT3050 et 3050FAVORIT), de codes couleurs (ADP2960WHM et ADP260WH)... Pour ces références, nous avons dû rechercher manuellement les correspondances. Des algorithmes de reconnaissance de mots existent mais sont difficiles à mettre en œuvre et peu efficaces sur des références.

Voici le résultat de la jointure des modèles entre IPC et GFK :

	Télévisions	Lave-vaisselle
Correspondances automatiques	388	223
Correspondances manuelles	168	104
Sans correspondance	33	11
Total des modèles IPC	589	338

Les modèles sans correspondance sont des modèles de marque SINGER, ou des modèles dont la référence est absente des fichiers GFK. On peut dans ce dernier cas imaginer que des orthographes trop différentes ont empêché un repérage manuel, ou bien penser qu'il existe un défaut d'exhaustivité GFK.

2.2.3. Agrégation des données IPC

La seconde tâche préalable à toute comparaison était de transformer les données IPC en vue d'obtenir un niveau d'agrégation identique à celui des données GFK. Ceci permet notamment de réaliser des calculs d'indices comparables à partir de deux sources différentes. Nous avons donc agrégé les séries élémentaires de l'IPC de façon à ce qu'elles constituent des séries au sens GFK vu

précédemment. Cette agrégation induit deux difficultés : le calcul des pondérations et le calcul des prix moyens.

a) Calcul des pondérations

Soit i une série de l'IPC (une marque-référence dans une forme de vente). $i \in \{1..S\}$

Soit j une agglomération. $j \in \{1..a\}$

A la série i correspondent plusieurs séries élémentaires. Notons k_{ij} l'une d'entre elles, observée dans l'agglomération j . $k_{ij} \in \{1..n_{ij}\}$

Soit N_j le nombre de séries élémentaires dans l'agglomération j . $N_j = \sum_{i=1}^S n_{ij}$

Le poids de la série i s'obtient par sommation des poids des séries élémentaires k_{ij} .

Or, au niveau de la var-agglo j , le poids (en valeur) d'une série élémentaire est de $\frac{1}{N_j}$. Au niveau de

la variété, le poids (en valeur) de la série élémentaire k_{ij} vaut donc :

$$pond(k_{ij}) = \frac{1}{N_j} pond(j) \quad \text{où } pond(j) \text{ est le poids (en valeur) de la var-agglo } j$$

Cette pondération est valable un an (de décembre à décembre) et peut évoluer d'une année à l'autre si l'échantillon est modifié (variation du nombre de séries élémentaires dans une ou plusieurs agglomérations par exemple). En cas de remplacement d'une série élémentaire, la série élémentaire remplaçante se trouve toujours dans la même agglomération et se voit affecter le poids de la série élémentaire remplacée. Le poids de la série élémentaire remplacée devient alors nul.

On en déduit le poids de la série i dans l'agglomération j :

$$pond(ij) = \sum_{k_{ij}=1}^{n_{ij}} \frac{1}{N_j} pond(j) = \frac{n_{ij}}{N_j} pond(j)$$

Et finalement, on obtient le poids de la série i :

$$pond(i) = \sum_{j=1}^a \frac{n_{ij}}{N_j} pond(j)$$

b) Calcul des prix moyens

Puisque nous disposons de pondérations en valeur, le calcul des prix moyens sur les données de l'IPC se fait à l'aide de moyennes harmoniques. La raison est simple :

$$ca = px * qt$$

$$pmoy = \frac{\sum_{i=1}^n px * qt}{\sum_{i=1}^n qt} = \frac{ca}{\sum_{i=1}^n \frac{ca}{px}}$$

Ce petit rappel effectué, voici comment nous avons calculé le prix moyen d'une série. En reprenant les notations précédentes et en se fixant un mois d'observation, on calcule le prix moyen de la série i dans l'agglomération j que l'on note $pmoy_{ij}$ à partir des prix des séries élémentaires $p_{k_{ij}}$.

$$pmoy_{ij} = \frac{n_{ij}}{\sum_{k=1}^{n_{ij}} \frac{1}{p_{k_{ij}}}}$$

On en déduit ensuite le prix moyen de la série i , noté $pmoy_i$.

$$p_{moy_i} = \frac{\sum_{j=1}^a \frac{n_{ij}}{N_j} pond(j)}{\sum_{j=1}^a \frac{n_{ij}}{N_j} \frac{pond(j)}{p_{k_{ij}}}}$$

2.2.4. Calcul d'indices comparés

Avec les pondérations et les prix moyens des séries i , $pond(i)$ et p_{moy_i} , on peut construire des indices de Laspeyres à niveau d'agrégation analogue, notés ILS_IPC et ILS_GFK, selon la formule :

$$ILS = \sum_{i=1}^s pond(i) I_i \quad \text{où } I_i = \frac{p_{moy_i}^1}{p_{moy_i}^0}$$

Ces deux indices sont assez différents de l'indice officiel de variété IPC, qui utilise des moyennes géométriques dans une de ces phases d'agrégation et dont le niveau de calcul élémentaire est plus fin. Il est à noter que ILS_IPC n'est pas un indice de Laspeyres sur séries élémentaires.

ILS_IPC et ILS_GFK possèdent en outre des différences de nature qui nuisent à leur comparabilité. Dans le cas des données IPC, $pond(i)$ et p_{moy_i} sont calculés d'après les méthodes exposées en 2.2.3. Dans le cas des données GFK, ils sont obtenus directement à partir des observations de chiffres d'affaires et de prix moyens des séries. En faisant intervenir des prix moyens par forme de vente, toutes agglomérations confondues, ILS_GFK reflète les effets de substitution entre agglomérations et entre points de vente¹². Dans le cas d'ILS_IPC en revanche, seuls des mouvements de séries élémentaires peuvent faire varier $pond(i)$ et p_{moy_i} . Les substitutions entre points de vente ne seront captées que lorsque des séries élémentaires d'un point de vente seront remplacées dans un autre point de vente, situation peu fréquente qui se produit surtout quand le point de vente initial ferme. De plus, des contraintes globales de nombre de séries à enquêter par agglomération et de répartition par forme de vente, limitent les transferts de relevés entre séries.

Des considérations qui précèdent (moindre prise en compte des substitutions par l'IPC), on s'attend donc à ce que, toutes choses égales par ailleurs, ILS_GFK soit inférieur à ILS_IPC. Par contre, la comparaison entre l'IPC officiel et ILS_IPC est moins aisée. Ces deux indices captent le phénomène de substitution, mais de façon différente : le premier via la moyenne géométrique, le second par la présence de prix moyens pondérés. Nous ne savons pas à ce stade hiérarchiser l'impact de ces deux méthodes.

¹² Le fait que l'on raisonne par forme de vente limite cependant l'effet des substitutions entre points de vente, puisqu'il ne s'agit que de substitutions opérées entre points de vente de la même forme de vente.

3. EVALUATION DE L'ECHANTILLON IPC A L'AIDE DES DONNEES GFK

3.1. Données de cadrage

TELEVISION	GFK	IPC
nombre de séries élémentaires	---	4047
nombre de séries	7095	1106
nombre de modèles	3348	589
nombre de marques	145	28

LAVE-VAISSELLE	GFK	IPC
nombre de séries élémentaires	---	1899
nombre de séries	4931	570
nombre de modèles	2085	338
nombre de marques	74	24

	TELEVISION		LAVE-VAISSELLE	
	GFK	IPC	GFK	IPC
nombre moyen de séries élémentaires par série	---	3,66	---	3,33
nombre moyen de séries par modèles	2,12	1,88	2,36	1,69
nombre moyen de modèles par marque	23,09	21,04	28,18	14,08

3.1.1. Pondérations

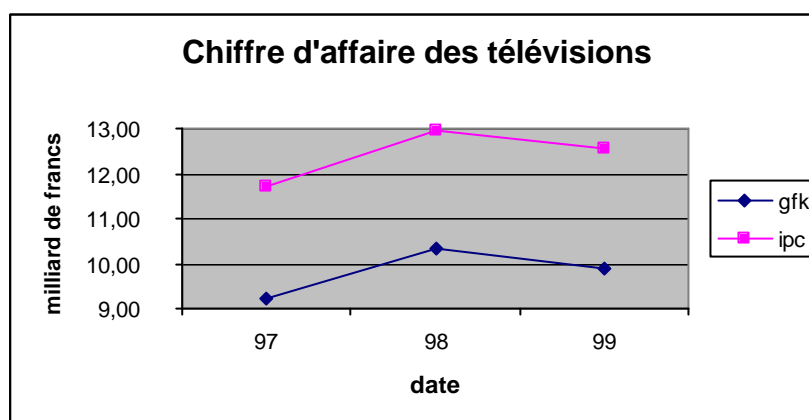
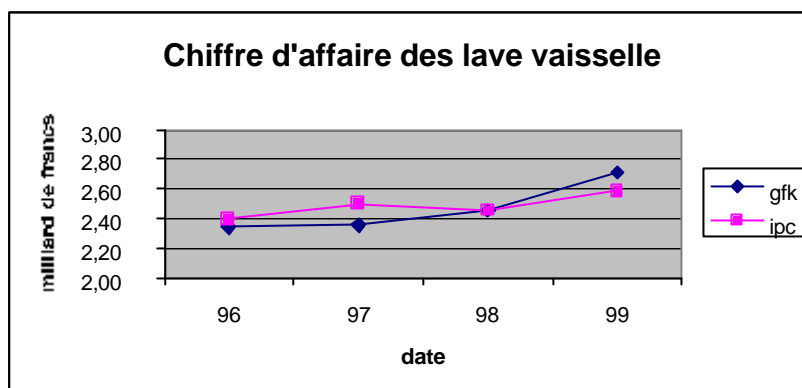
Les pondérations IPC sont normalisées. Or, nous disposons de l'information permettant de calculer à partir de ces pondérations les montants de consommation associés. Nous les noterons *ca* (pour chiffre d'affaires). Ces montants de consommation seront alors comparables aux montants GFK obtenus en multipliant les prix et les quantités.

Cependant se pose au préalable un problème de définition d'année. Il est difficile de reconstituer des chiffres d'affaires annuels à partir des données GFK car le bimestre d'observation décembre-janvier appartient à deux années consécutives. Nous avons donc redéfini la notion d'année : la nouvelle année *n* s'étend de février de l'année *n* à janvier de l'année *n+1*. Cette décomposition présente l'intérêt de maximiser le nombre d'années observées puisque les données GFK s'étendent de février 97 pour les télévisions (de février 96 pour les lave-vaisselle) à janvier 2001. On obtient ainsi 4 années entières d'observation pour les téléviseurs, et 5 pour les lave-vaisselle.

Les montants de consommation GFK sont faciles à calculer, il suffit des sommer les montants des 6 bimestres de chaque nouvelle année. En revanche il faut adapter les montants annuels calendaires de l'IPC. Ici la nouvelle année comprend 11 mois de la même année et 1 mois de l'année suivante. Puisqu'au sein d'une même année calendaire les pondérations sont fixes, il est légitime de faire une moyenne prorata temporis des chiffres d'affaires des deux années (calendaires). Ainsi, en prenant l'exemple de l'année 98, on calcule le nouveau montant de consommation IPC de la manière suivante :

$$ca_{98\text{nouveau}} = \frac{11}{12} ca_{98\text{calendaire}} + \frac{1}{12} ca_{99\text{calendaire}}$$

Les chiffres d'affaires annuels des deux produits sont représentés pour l'IPC et GFK sur les deux graphiques ci-après.



Pour les lave-vaisselle, l'écart entre les chiffres d'affaires IPC et GfK est faible. Il oscille entre -4,5% (en 99) et +6% (en 97). Pour les télévisions, l'écart est plus important mais oscille peu : entre 25% et 27%. Ces écarts peuvent paraître élevé mais il faut les replacer dans leur contexte, soit entre 4000 et 4500 milliards de francs en montant de consommation totale couvert par l'IPC chaque année. L'écart maximum observé (environ 2,6 milliards en 98 pour les télévisions) ne représente que 0,06% de la consommation totale couverte de l'année.

On notera que pour les deux produits, le profil des séries est le même selon les deux sources.

Dans la suite, nous ventilerons le chiffre d'affaires cumulé sur la globalité de la période d'observation en fonction de différents critères (forme de vente, marque, modèle, CTs) afin d'apprécier la qualité de l'échantillon IPC.

Chiffres d'affaires cumulés en milliards de francs

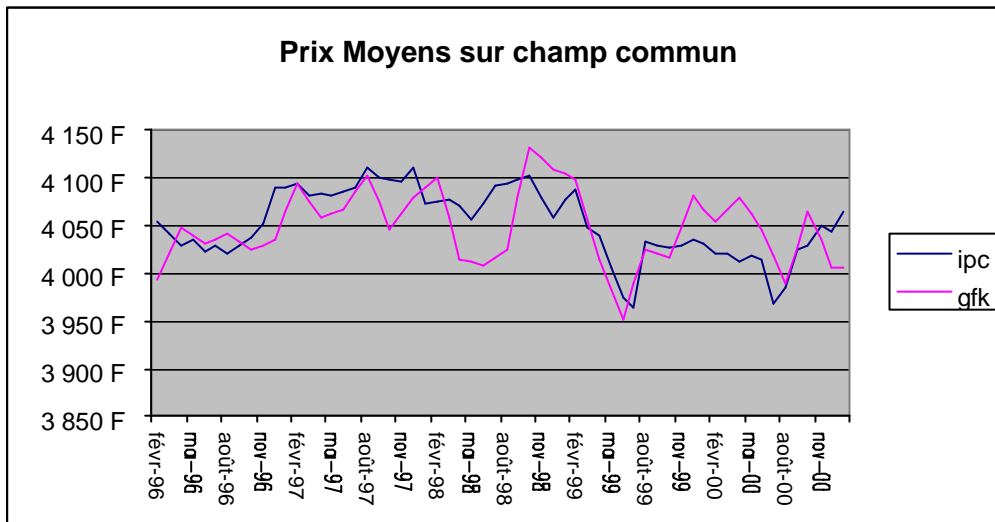
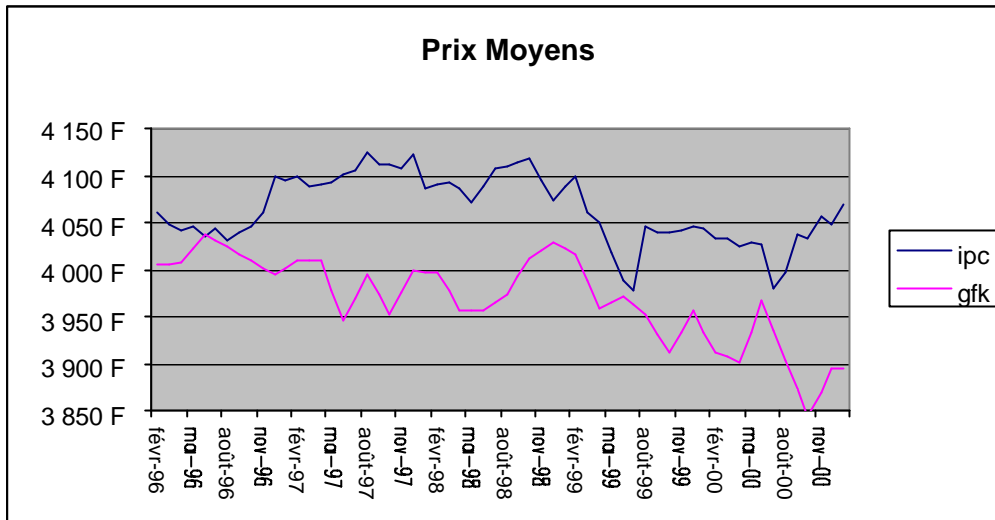
	<i>lave-vaisselle</i>	<i>télévisions</i>
gfk	9,86	29,45
ipc	9,93	37,20

3.1.2. Prix moyens

Conformément à la partie 2, nous avons calculé, à chaque date d'observation, des prix moyens GfK par moyenne arithmétique pondérée par les quantités, et des prix moyens IPC par moyenne harmonique pondérée par les montants de consommation.

Les tableaux suivants présentent la chronologie de ces prix moyens, à la fois sur champ total (c'est à dire sur l'ensemble des données) et sur champ commun (ie sur l'ensemble des modèles communs, voir explication au paragraphe 2.2.2.). Pour cela nous avons réalisé une interpolation linéaire des prix moyens GfK entre 2 bimestres d'observation consécutifs.

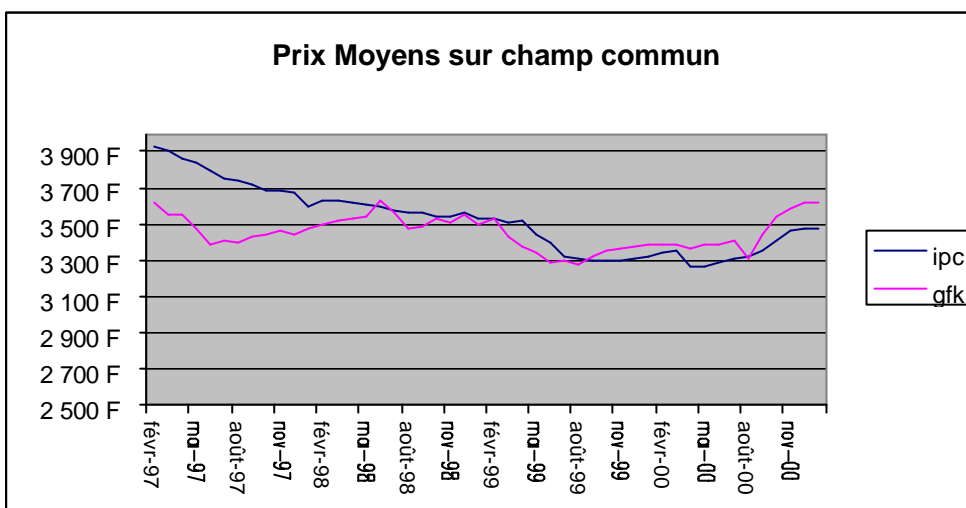
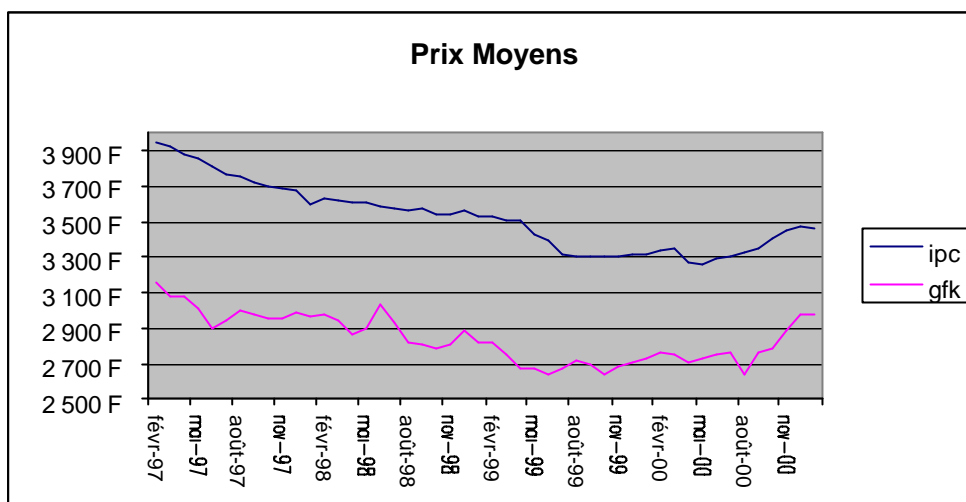
cas des lave -vaisselle



Prix moyens globaux

	IPC	GFK	ECART
totalité des fichiers	4 064,26 F	3 966,61 F	2,46%
champ commun	4054,11 F	4 047,77 F	0,16%

cas des télévisions



Prix moyens globaux

	IPC	GFK	ECART
totalité des fichiers	3 501,93 F	2 841,92	23,22%
champ commun	3 503,47 F	3 452,68	1,47%

La comparaison sur champ total révèle une surestimation systématique des prix moyens par l'IPC, surtout pour les téléviseurs, tandis que la comparaison sur champ commun génère des courbes très proches, en particulier pour les lave-vaisselle où l'écart bimestriel ne dépasse pas 200F.

Si les prix moyens de notre échantillon sont plus élevés, la raison est donc à rechercher du côté du champ de collecte. Nous reviendrons sur cette notion par la suite, mais on peut déjà deviner dans l'IPC une tendance à la sélection des modèles de standing supérieur, que cette supériorité relève de performances techniques ou bien de l'image de marque. Si l'on considère que les marques référencées par GFK approximent l'existant des marques, il apparaît d'ailleurs que les marques IPC correspondent aux marques les plus connues de l'existant, ie les marques de haut standing. On peut comprendre alors que la courbe IPC soit toujours située au dessus de la courbe GFK, et qu'elle le soit davantage pour les téléviseurs où la proportion de marques communes est de 19%, contre 31% pour les lave-vaisselle.

3.2. Répartition par forme de vente

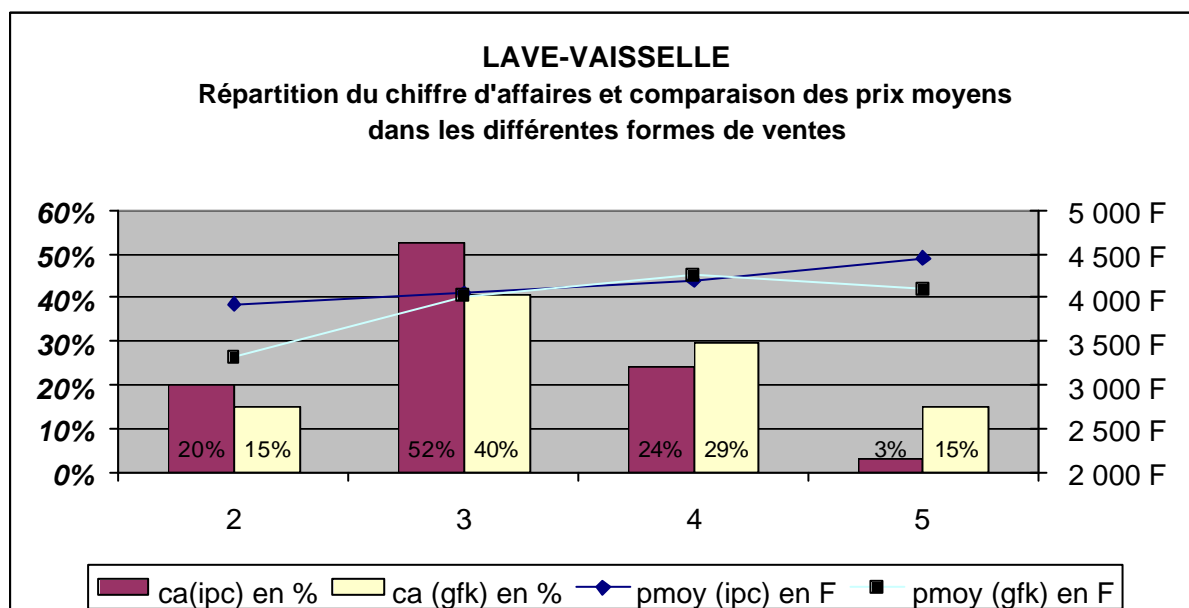
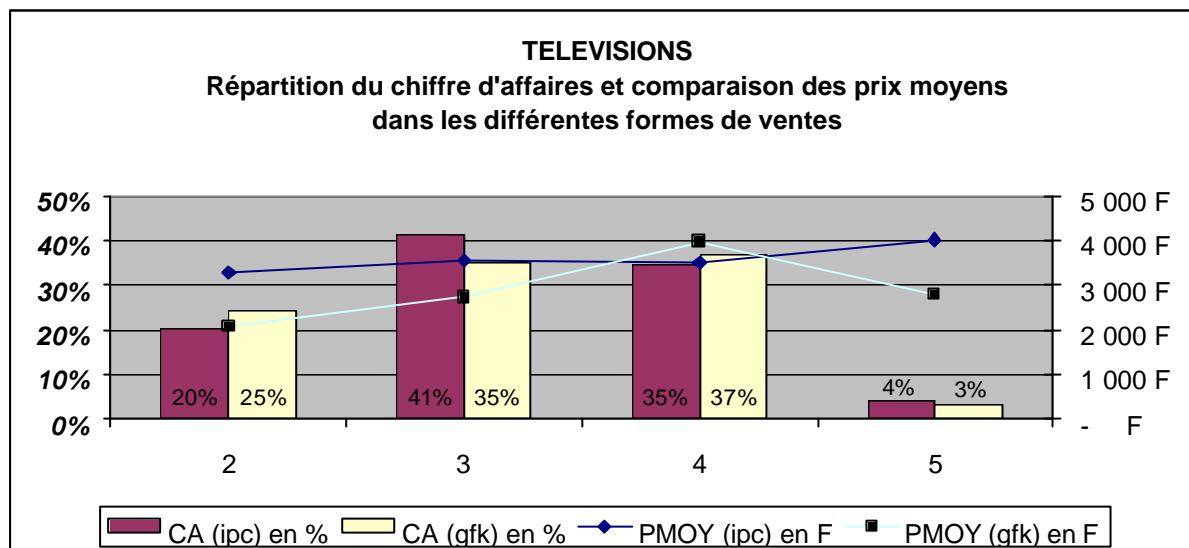
Les tableaux ci-dessous présentent la répartition des chiffres d'affaires IPC et GFK selon la variable *forme de vente*, dont nous rappelons les modalités :

'2' = *Hypermarchés*

'3' = *GMS*

'4' = *Magasins traditionnels*

'5' = *Autres (Grands magasins, VPC¹³)*.



On voit déjà, d'après ces graphiques, que la répartition par forme de vente choisie par l'IPC « colle » assez bien à la réalité du marché : la distribution des chiffres d'affaires selon les différents réseaux est proche de celle donnée par GFK, exception faite de la forme de vente « autres » pour les lave-vaisselle. L'explication vient des spécialistes de cuisine : ils représentent pour GFK une part importante de la forme de vente, alors que l'IPC possède très peu de points de vente de ce type dans son échantillon (au profit des magasins de meubles ou d'électroménager). Voilà pourquoi le

¹³ Pour l'IPC cette dernière modalité ne correspond qu'aux grands magasins, les relevés en VPC n'appartenant pas au champ de collecte « Biens Durables ». Pour GFK elle comprend en outre les spécialistes de cuisine.

pourcentage IPC associé à la forme de vente '5' est sensiblement inférieur au pourcentage GFK, tandis que la situation est inversée pour la forme de vente '3'.

Les observations sur prix moyens précisent les conclusions menées dans la section précédente. Là encore les prix moyens des lave-vaisselle sont plus proches que ceux des télévisions, du fait d'un marché moins volatil. Mais il est intéressant surtout de noter que les écarts que nous avons soulignés en 3.1.2. , et qui mettaient en évidence une tendance à la sélection des modèles de standing supérieur, concernent principalement les formes de vente '2' et '5'. En effet c'est dans les hypermarchés et dans la VPC que l'on trouve essentiellement les marques « bas de gamme », que GFK inclut dans ses relevés, mais que l'IPC sous-représente de fait.

3.3. Marque

Regardons maintenant ce qu'il en est de la répartition par marque des chiffres d'affaires et des prix moyens. Un classement décroissant des marques par chiffres d'affaires nous permet d'attribuer un « rang IPC » aux marques de l'IPC, et un « rang GFK » aux marques GFK. L'objectif des tableaux ci-dessous est de comparer les valeurs prises par ces deux variables de rang sur les marques de l'échantillon IPC. Ils sont classés suivant le « rang IPC », ce qui facilite la comparaison.

Télévisions	GFK				IPC		
	MARQUE	rang GFK	ca marque / ca total	ca marque / ca restreint	prix moyen	rang IPC	ca marque / ca total
PHILIPS	1	23,96%	26,88%	4 275 F	1	36,33%	3 661 F
SONY	2	16,97%	19,03%	4 469 F	2	25,16%	3 735 F
THOMSON	3	12,56%	14,09%	3 639 F	3	17,66%	3 683 F
GRUNDIG	4	4,90%	5,49%	2 918 F	4	5,44%	2 850 F
RADIOLA	7	3,10%	3,48%	2 299 F	5	4,46%	2 799 F
SCHNEIDER	8	2,88%	3,23%	1 809 F	6	2,38%	2 475 F
TOSHIBA	9	2,75%	3,08%	3 408 F	7	1,52%	3 035 F
PANASONIC	6	3,58%	4,02%	4 281 F	8	1,48%	3 527 F
BRANDT	5	4,59%	5,15%	1 959 F	9	0,95%	2 265 F
HITACHI	17	1,39%	1,56%	3 023 F	10	0,87%	2 984 F
SINGER					11	0,87%	3 460 F
SABA	11	1,95%	2,19%	2 368 F	12	0,72%	3 301 F
AKAI	23	0,55%	0,61%	2 631 F	13	0,49%	4 070 F
TELEFUNKEN	30	0,35%	0,39%	3 974 F	14	0,40%	4 935 F
LOEWE	18	0,93%	1,05%	8 689 F	15	0,20%	6 097 F
MITSUBISHI	21	0,58%	0,65%	2 451 F	16	0,17%	2 460 F
GOLDSTAR	12	1,89%	2,12%	2 119 F	17	0,17%	2 385 F
DESMET	26	0,47%	0,53%	1 371 F	18	0,13%	3 737 F
JVC	22	0,56%	0,63%	3 486 F	19	0,13%	3 715 F
SHARP	16	1,44%	1,61%	1 883 F	20	0,12%	2 423 F
SIEMENS	76	0,01%	0,01%	3 764 F	21	0,08%	3 667 F
DIGILINE	74	0,01%	0,02%	3 576 F	22	0,07%	6 282 F
ITTOCEANIC	48	0,07%	0,08%	3 874 F	23	0,07%	3 709 F
SAMSUNG	15	1,58%	1,77%	2 281 F	24	0,06%	2 995 F
DAEWOO	13	1,71%	1,92%	1 721 F	25	0,02%	1 354 F
SCHAUBLorenz	52	0,05%	0,06%	1 581 F	26	0,02%	3 302 F
MADISON	61	0,04%	0,04%	1 784 F	27	0,01%	3 290 F
FIRSTLINE	31	0,28%	0,31%	1 952 F	28	0,00%	1 746 F
TOTAL		89,15%	100,00%	3 248 F		100,00%	3502 F

Lave-vaisselle MARQUE	GFK				IPC		
	rang GFK	ca marque / ca total	ca marque / ca restreint	prix moyen	rang IPC	ca marque / ca total	prix moyen
BRANDT	2	11,46%	12,82%	3 828 F	1	25,40%	3 760 F
WHIRLPOOL	1	15,95%	17,85%	4 011 F	2	23,56%	4 041 F
THOMSON	7	5,04%	5,64%	4 169 F	3	8,45%	4 212 F
ARTHURMARTIN	5	7,88%	8,82%	3 964 F	4	8,19%	4 046 F
MIELE	3	9,22%	10,32%	6 355 F	5	7,82%	5 880 F
BOSCH	4	8,63%	9,66%	4 730 F	6	7,08%	4 434 F
VEDETTE	8	4,56%	5,10%	3 276 F	7	6,68%	3 544 F
SIEMENS	6	7,58%	8,48%	4 379 F	8	3,69%	4 270 F
SINGER					9	1,91%	4 831 F
AEG	9	2,74%	3,06%	4 817 F	10	1,84%	4 520 F
LADEN	15	1,84%	2,06%	2 994 F	11	1,34%	3 188 F
FAURE	13	2,06%	2,31%	3 065 F	12	0,97%	3 253 F
ZANUSSI	23	0,61%	0,68%	2 916 F	13	0,67%	2 998 F
CANDY	14	1,94%	2,17%	3 356 F	14	0,46%	3 241 F
INDESIT	12	2,16%	2,42%	2 698 F	15	0,39%	2 950 F
DEDIETRICH	19	1,40%	1,57%	4 603 F	16	0,37%	4 886 F
BAUKNECHT	18	1,61%	1,80%	4 814 F	17	0,33%	5 202 F
ARISTON	11	2,42%	2,71%	3 464 F	18	0,32%	3 468 F
SCHOLTES	16	1,72%	1,92%	4 198 F	19	0,26%	3 992 F
RADIOLA	33	0,17%	0,19%	3 095 F	20	0,20%	3 773 F
NOGAMATIC	73	0,00%	0,00%	1 990 F	21	0,04%	2 568 F
HOOVER	31	0,19%	0,21%	2 922 F	22	0,02%	3 490 F
ASPES	34	0,17%	0,18%	2 256 F	23	0,02%	2 390 F
SIDEX	65	0,00%	0,00%	2 398 F	24	0,01%	2 630 F
TOTAL		89,34%	100,00%	4 070 F		100,00%	4 064 F

Qu'il s'agisse des téléviseurs ou des lave-vaisselle, les 9 « premières » marques de l'échantillon IPC, représentant plus de 90% de son chiffre d'affaires, sont les 9 « premières » marques GFK (exception faite de la marque Singer). La correspondance des rangs au sein de ces 9 marques est excellente pour les téléviseurs. Elle est un peu moins bonne pour les lave-vaisselle, mais reste satisfaisante : en particulier la « hiérarchie » reste respectée ; Brandt et Whirlpool conservent les deux premières positions malgré leur interversion. On remarque que la part de chiffre d'affaires associée aux quatre premières marques est toujours supérieure aux proportions qu'affichent les données GFK. Ceci rejoint l'idée déjà énoncée d'une sélection orientée vers les marques les plus connues, mais peut aussi venir de ce que le nombre de marques suivies dans l'échantillon IPC est plus réduit (5 fois moins que dans les données GFK pour les téléviseurs, 3 fois moins pour les lave-vaisselle).

Au delà du dixième rang, on retrouve parmi les marques de l'IPC des marques dont le rang GFK varie de 11 à 76. L'échantillon IPC se répartit donc uniformément entre les marques restantes, sans forcément retenir les chiffres d'affaires les plus importants, ni prendre en compte les marques situées en fin de classement. Ce second phénomène est totalement justifié, les derniers rangs correspondant à des marques marginales, dont la durée d'observation des séries ne dépasse pas le bimestre dans les fichiers GFK. Notre Indice de Prix semble donc retenir d'office les marque leader, puis les compléter par un échantillon de marques moins connues. Il en résulte une image correcte du marché, déformée toutefois au profit des marques leader qui sont sur-représentées.

Il peut paraître injustifié cependant de ne pas voir apparaître au rang des marques suivies certaines marques « bien classées » par GFK, par exemple la marque Bluesky (rang 14 sur 145) : il s'agit d'une marque nouvelle, ayant réussi à pénétrer le marché depuis peu via des prix compétitifs en hypermarchés, et dont l'IPC n'a peut-être pas encore perçu la progression. Il aurait été intéressant d'ailleurs d'effectuer une analyse temporelle de cette répartition des marques, pour repérer à quelle

vitesse l'IPC intègre les évolutions structurelles (marques gagnant ou perdant des parts de marché) mises en évidence par la variable « rang GFK ».

Enfin en ce qui concerne les prix, on peut noter que le prix moyen des grandes marques a tendance à être plus faible dans l'IPC que dans GFK, le phénomène s'atténuant (voire s'inversant) sur les autres populations de marques. Cette observation sera analysée dans la section suivante, à la lumière de nouveaux résultats.

3.4. Modèle

Comment les modèles suivis dans l'IPC sont-ils choisis ? Prend-on les modèles les plus vendus, les moins vendus ? Ou équilibre-t-on l'échantillon des modèles ?

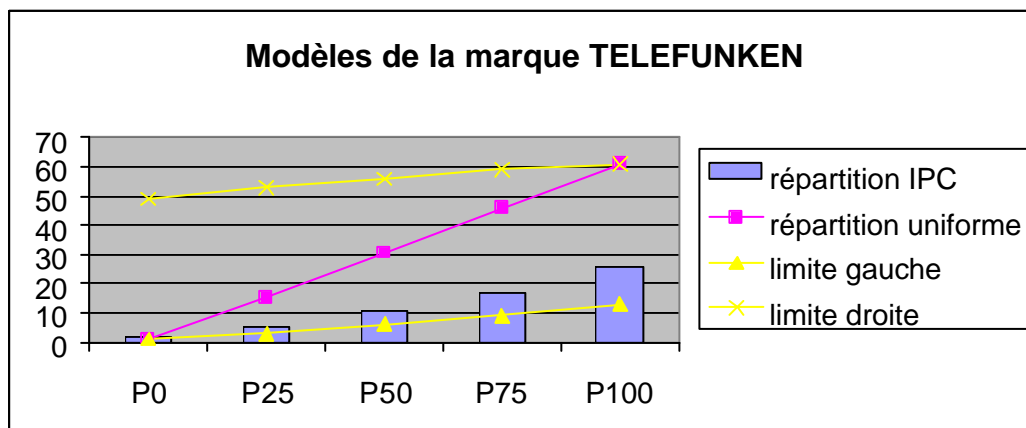
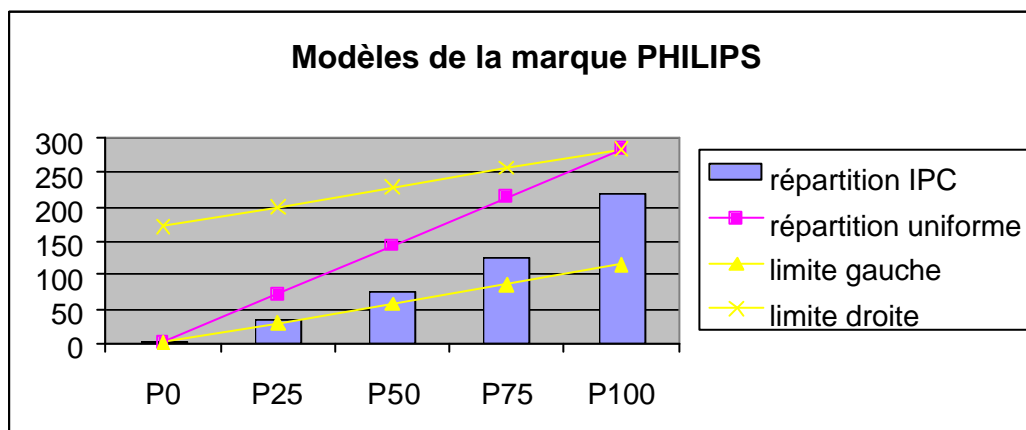
Afin de mieux appréhender cette question, nous avons classé, dans les données GFK, les modèles d'une marque par chiffre d'affaires décroissant en leur attribuant un rang. Il suffit d'étudier la distribution de ce rang dans l'échantillon des modèles IPC suivis. Si cette distribution est concentrée à gauche, les modèles sélectionnés sont les modèles les mieux vendus. Si la distribution est concentrée à droite, ce sont les modèles les moins vendus qui ont été sélectionnés. Enfin, si la distribution est uniforme, le choix des modèles est équilibré.

Les tableaux ci dessous représentent, pour chaque marque de l'échantillon IPC, le nombre de modèles GFK, le nombre de modèles IPC, ainsi que le nombre de modèles IPC avec correspondance GFK (voir partie 2.2.2.). Pour apprécier la distribution dans l'IPC du rang des modèles sélectionnés, nous disposons du minimum (P0), du premier et du troisième quartile (P25 et P75), de la médiane (P50) et du maximum (P100). Ainsi, par exemple, 25% (P25) des 36 modèles IPC Whirlpool communs avec GFK sont sélectionnés parmi les 10 premiers modèles vendus selon GFK.

Lave-vaisselle MARQUE	GFK		IPC							
	rang GFK	nombre de modèles	rang IPC	nombre de modèles	nombre de modèles communs	distribution des rangs des modèles				
						P0	P25	P50	P75	P100
BRANDT	2	130	1	40	40	1	10,5	29,5	57	119
WHIRLPOOL	1	179	2	37	36	1	10	29,5	80,5	127
THOMSON	7	86	3	36	36	1	10,5	25	37,5	61
ARTHURMARTIN	5	103	4	25	25	1	7	23	41	88
MIELE	3	91	5	20	20	1	4,5	9,5	21,5	90
BOSCH	4	169	6	35	35	2	16	33	53	94
VEDETTE	8	92	7	36	36	1	9,5	24	49	82
SIEMENS	6	209	8	25	24	1	7	14,5	48	208
SINGER			9	4	0					
AEG	9	72	10	13	13	1	6	13	19	53
LADEN	15		11	13	12	1	4,5	9,5	15	23
FAURE	13	53	12	14	14	2	8	18,5	26	46
ZANUSSI	23	33	13	4	4	2	2,5	4,5	6,5	7
CANDY	14	43	14	5	5	2	6	8	11	20
INDESIT	12	18	15	2	2	1	1	2	3	3
DEDIETRICH	19	58	16	4	3	1	1	29	36	36
BAUKNECHT	18	67	17	8	8	1	5,5	16	20,5	38
ARISTON	11	69	18	5	5	4	4	14	18	30
SCHOLTES	16	66	19	3	3	1	1	5	13	13
RADIOLA	33	14	20	5	3	7	7	8	9	9
NOGAMATIC	73	1	21	1	0					
HOOVER	31	17	22	1	1	8	8	8	8	8
ASPES	34	5	23	1	1	2	2	2	2	2
SIDEX	65	5	24	1	1	3	3	3	3	3

Télévisions MARQUE	GFK		IPC							
	rang GFK	nombre de modèles	rang IPC	nombre de modèles	nombre de modèles communs	distribution des rangs des modèles				
						P0	P25	P50	P75	P100
PHILIPS	1	285	1	120	115	1	34	74	124	220
SONY	2	193	2	79	78	1	19	44	76	135
THOMSON	3	285	3	104	100	1	26	64,5	116	275
GRUNDIG	4	209	4	61	59	1	17	46	74	134
RADIOLA	7	101	5	32	32	1	7,5	17	36	87
SCHNEIDER	8	70	6	16	15	1	4	10	19	48
TOSHIBA	9	82	7	19	19	1	7	16	27	64
PANASONIC	6	129	8	34	34	2	20	33	68	111
BRANDT	5	66	9	21	19	1	5	16	24	66
HITACHI	17	77	10	16	16	1	5,5	12,5	23	65
SINGER			11	10	0					
SABA	11	86	12	13	11	1	14	22	26	29
AKAI	23	39	13	7	7	2	7	14	22	23
TELEFUNKEN	30	61	14	12	10	2	5	11	17	26
LOEWE	18	93	15	3	3	6	6	9	10	10
MITSUBISHI	21	47	16	4	4	1	3,5	7	10	12
GOLDSTAR	12	75	17	6	6	1	4	14,5	25	29
DESMET	26	77	18	5	2	7	7	13,5	20	20
JVC	22	41	19	3	3	1	1	12	20	20
SHARP	16	56	20	5	5	2	7	9	12	13
SIEMENS	76	14	21	4	4	1	2	5,5	9	10
DIGILINE	74	5	22	1	0					
ITTOCEANIC	48	56	23	5	5	2	6	8	23	39
SAMSUNG	15	70	24	4	4	2	4	16	34	42
DAEWOO	13	50	25	2	2	16	16	23	30	30
SCHAUBLORENZ	52	18	26	1	1	3	3	3	3	3
MADISON	61	6	27	1	1	1	1	1	1	1
FIRSTLINE	31	35	28	1	1	31	31	31	31	31

Les graphiques représentent, pour une marque donnée, la distribution des rangs des modèles (répartition IPC), ainsi que des repères montrant les distributions particulières : limite gauche (modèles les plus vendus), limite droite (modèles les moins vendus) et distribution uniforme.



Les tableaux et graphiques précédents permettent de répondre aux questions énoncées en tout début de section. Ils révèlent une **concentration à gauche** de la distribution des modèles enquêtés, aussi bien pour les lave-vaisselle que pour les téléviseurs. Cependant un **facteur marque** vient s'ajouter : la concentration est plus forte pour les marques de moyenne et basse gamme. L'exemple des marques Philips et Téléfunken est révélateur : on observe jusqu'au 220^{ème} modèle Philips (sur 285) tandis que l'on s'arrête au 26^{ème} modèle Téléfunken (sur 61).

Ainsi la règle du « bien suivi, bien vendu » déséquilibre l'échantillon des modèles IPC, en orientant le choix des enquêteurs vers les modèles à plus fort chiffre d'affaires. Plus la renommée de la marque est faible, plus la déviation s'accroît, puisqu'alors seuls quelques modèles (les plus perfectionnés généralement) sont effectivement bien vendus.

Ce phénomène précise les conclusions que nous avons menées précédemment. La sélection de modèles « bien vendus » conduit à retenir ou bien des modèles de grande marque, ou bien des modèles hauts de gamme parmi les marques modestes. Dans les deux cas de figure, sont sélectionnés des modèles de standing supérieur, par conséquent des modèles chers.

On peut également essayer d'expliquer maintenant les différences de prix moyens par marque mentionnés en 3.3. Pour les grandes marques, les modèles les plus vendus correspondent aux modèles « médians » du point de vue des caractéristiques techniques. En favorisant l'observation de ces modèles, l'IPC éliminerait de son champ de collecte les articles de très haute gamme, ce qui générerait des prix moyens inférieurs aux prix GFK. Pour les marques bas de gamme, la configuration inverse serait due à une sélection par les enquêteurs des modèles les plus évolués. Ces hypothèses demandent à être vérifiées, mais l'exemple des marques Philips (prix moyen IPC de 3661F, pour un prix moyen GFK de 4275F) et Téléfunken (prix moyen IPC de 4935F, pour un prix GFK de 3974F) les conforte.

3.5. Caractéristiques techniques

Nous allons maintenant aborder un dernier mode d'évaluation de la qualité de notre Indice de Prix. Il s'agit de comparer les répartitions par chiffres d'affaires avec celles des données GFK, suivant quelques variables de caractérisation nous ayant paru pertinentes.

3.5.1. Cas des lave-vaisselle

L'IPC relève quasi-exclusivement des lave-vaisselle de type « pose libre » (posable). Ce tableau révèle pourtant que plus de 22% du chiffre d'affaires GFK est attribué aux lave-vaisselle intégrables, dont le prix moyen est plus élevé.

	nb de série	ca	%ca	prix moyen
encastrable	87	87 385 235 F	0,69%	3 049 F
posable	2863	9 768 952 467 F	76,96%	3 822 F
intégrable	1981	2 837 151 889 F	22,35%	4 612 F
Total	4931	12 693 489 591 F	100,00%	

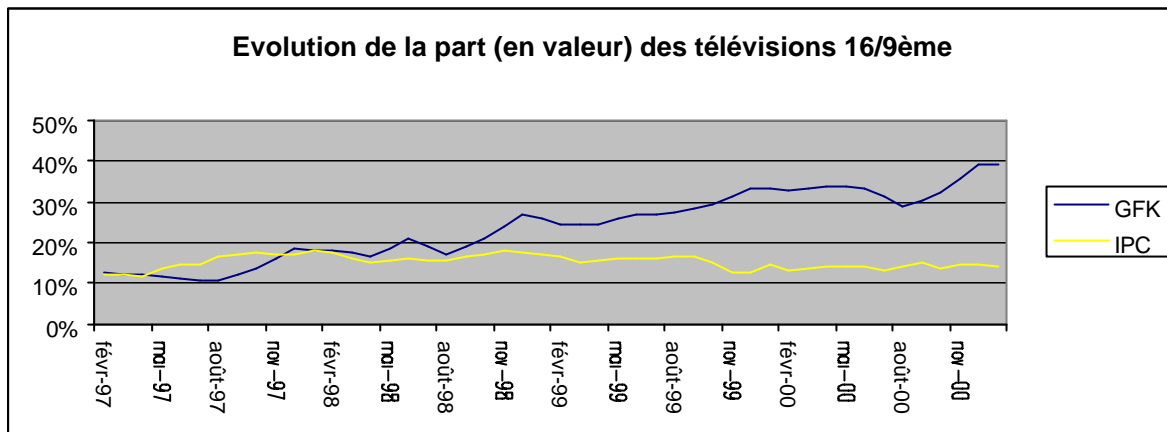
3.5.2. Téléviseurs : taille d'écran

taille d'écran	ca (ipc)	ca (gfk)	pmoy (ipc)	pmoy (gfk)
11-25 cm	0%	1%		1 895 F
36-50 cm	3%	12%	1 488 F	1 236 F
51-56 cm	29%	21%	2 663 F	1 931 F
60-68 cm	21%	8%	4 014 F	4 148 F
70-76 cm	37%	43%	4 682 F	4 135 F
81-87 cm	1%	13%	9 308 F	9 066 F
92-132 cm	0%	1%		19 699 F
manquant	9%	0%	3 308 F	

L'étude de la répartition par taille d'écran est également intéressante. Si l'on comprend que les appareils dont la diagonale est supérieure à 92 cm (rétroprojecteurs) ou inférieure à 25 cm (écrans à cristaux liquides) ne soient pas relevés, les proportions de téléviseurs de classe 2 ('36-50 cm') et 6

('81-87cm') sont inadaptées. Les collecteurs de prix étant tenus de choisir des produits conformes à la définition de la variété (ici : *récepteur TV couleur, exclure combinés, diagonale >=36 cm*), on peut avancer comme explication un « effet de seuil » les incitant à enquêter des articles dont la taille se situe après la borne autorisée. Un autre facteur tient à ce que l'on vient de voir précédemment : la règle du « bien suivi, bien vendu » gonfle la part des '51-68 cm' au détriment des autres classes de diagonales.

3.5.3. Téléviseurs : format

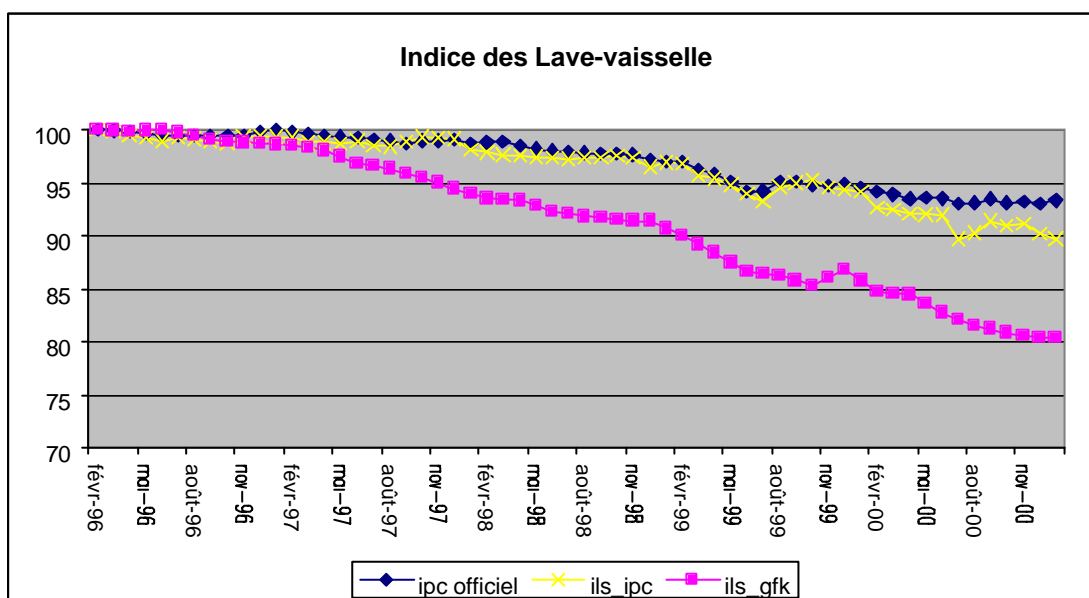
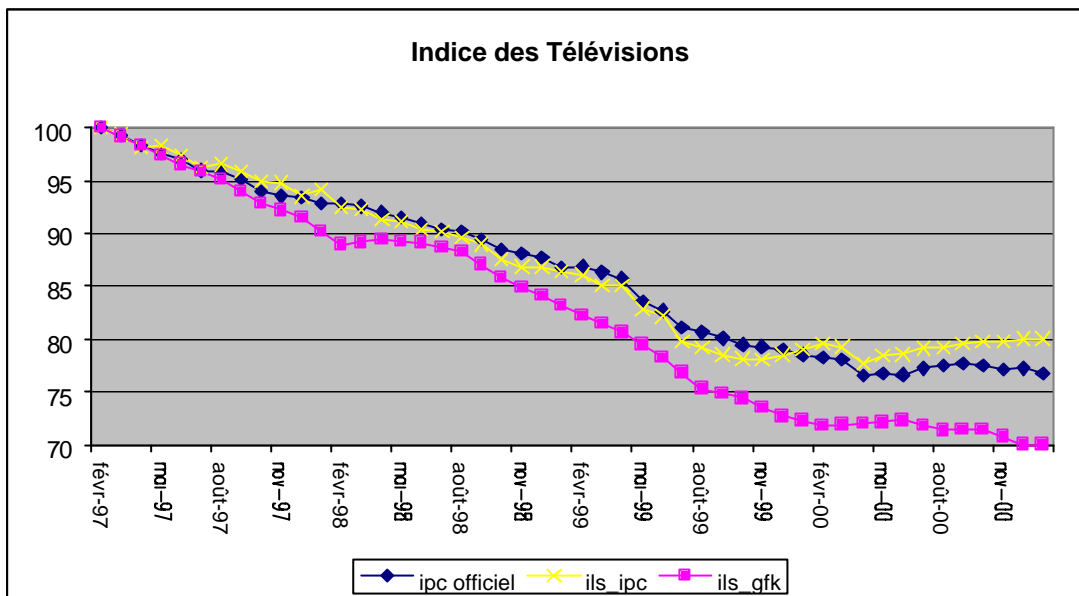


Ce graphique met en évidence une stabilité du chiffre d'affaires associé aux téléviseurs 16/9^{ème} de notre échantillon, alors que les estimations des données GFK « décollent » à partir de mi-98. Autrement dit, l'IPC n'a pas encore pris en compte l'essor des téléviseurs 16/9^{ème} sur le marché.

3.6. Indices

Pour chaque produit, on calcule trois indices :

- ?? Le premier indice (IPC officiel) est l'indice de variété IPC. La population considérée est la totalité des séries élémentaires. Le calcul de cet indice se décompose suivant deux formules : la première est une moyenne géométrique équipondérée des micro-indices, la seconde un Laspeyres des indices de var-agglos. De plus, il prend en compte le traitement de la qualité lors des remplacements de produits.
- ?? Le deuxième indice (ILS_IPC) est construit à partir des données IPC agrégées (ie les séries, voir partie 2.2.4.). C'est un indice de Laspeyres calculé en base 100 en février de chaque année (c'est le début de l'année dans sa nouvelle acception) puis chaîné annuellement pour obtenir finalement des indices en base 100 en février 96 pour les lave-vaisselle, et en février 97 pour les téléviseurs. Cet indice ne prend en compte aucun traitement de qualité (ce n'est pas l'objet de l'étude) ce qui nous contraint à restreindre le champ des séries en ne considérant que les séries présentes sans interruption de février à février de l'année suivante.
- ?? Le troisième indice (ILS_GFK) est quant à lui construit à partir des séries GFK sur le même principe que le deuxième indice. La fréquence des observations étant bimestrielle, nous avons mensualisé cet indice par simple interpolation linéaire.



La situation est similaire pour les deux produits. L'IPC officiel est étonnamment proche de l'ILS_IPC, malgré les différences existant entre ces deux indices. Il n'y a que 4.2% d'écart au bout de 4 années d'observation pour les télévisions, et 3.9% pour les lave-vaisselle au bout de 5 années.

ILS_GFK par contre a une tendance à la baisse nettement plus prononcée, en particulier pour les lave-vaisselle, avec un écart de l'ordre de 8-10 points par rapport à l'ILS_IPC. C'est pourquoi nos analyses se sont portées sur ce seul produit, les conclusions obtenues étant transposables au cas des téléviseurs.

Trois facteurs peuvent contribuer à expliquer de tels écarts d'évolution : la population prise en compte, ses prix et ses pondérations. Nous avons donc simulé 4 formules d'indices, différenciées par la fréquence de chaînage et le mode de pondération, sur 3 champs (champ total IPC, champ total GFK et champ commun). L'objectif ici n'étant pas d'aborder les ajustements de qualité, nous nous sommes restreints aux populations présentes (sans discontinuité) au sein d'une période de chaînage. On comprend alors aisément que plus la période de chaînage est longue, plus la population sélectionnée est réduite.

Pourcentage moyen de chiffre d'affaires par période pour les séries sélectionnées

	non chaîné	chaînage annuel	chaînage bimestriel
IPC (champ total)	1,1%	45,4%	96,5%
IPC (champ commun)	1,1%	44,7%	96,5%
GFK (champ total)	2,0%	56,7%	92,1%
GFK (champ commun)	1,7%	62,3%	96,1%

Le calcul d'indices non chaînés, sur des séries présentes tout au long de la période, n'a pas de sens du fait d'effectifs trop faibles. Nous nous sommes donc limités aux cas des fréquences de chaînage annuelle et bimestrielle.

Indices en janvier 2001

	champ total			champ commun		
	ILS_IPC	ILS_GFK	écart brut	ILS_IPC	ILS_GFK	écart brut
chaînage annuel pondéré	89,75	80,40	9,34	89,49	78,99	10,50
<i>chaînage annuel non pondéré</i>	93,96	88,08	5,88	93,75	83,61	10,14
chaînage bimestriel pondéré	88,85	84,25	4,60	88,62	83,40	5,22
<i>chaînage bimestriel non pondéré</i>	91,34	96,36	-5,02	91,08	89,49	1,59

Il est logique d'obtenir des résultats similaires sur les deux champs pour l'ILS_IPC, puisque ces derniers sont quasi identiques (voir partie 2.2.2.). Les ILS_GFK pondérés sont également proches d'un champ à l'autre : ceci nous prouve que parmi l'ensemble des modèles GFK, ce sont ceux sélectionnés par l'IPC qui «font» l'indice. Se restreindre au champ commun ne rapproche pas les indices : l'écart entre ILS_IPC et ILS_GFK a même tendance à légèrement augmenter.

Si l'on s'intéresse aux indices GFK, on remarque que les indices non pondérés sont plus hauts que les indices pondérés, et que la hausse est favorisée par les facteurs 'champ total' et 'chaînage bimestriel'. Ce phénomène est dû à l'existence de « petits modèles » (ie faiblement pondérés), dont les diminutions de prix sont moindres. Ces modèles, contenus dans le champ total GFK mais absents du champ commun, ont des durées de vie généralement courtes, ce qui explique que leur influence soit accentuée par un chaînage bimestriel. Les « petits modèles » tirent donc vers le haut ILS_GFK lorsqu'on passe du champ commun au champ total, alors que l'effet est bien sûr beaucoup plus limité pour l'indice ILS_IPC, ce qui explique que la restriction au champ commun ne réduise pas les écarts.

Approfondissons maintenant l'analyse sur les données pondérées du champ commun. Lorsqu'on passe du chaînage annuel au chaînage bimestriel, l'écart se réduit de moitié (de 10,5 à 5,2 points). Si l'ILS_IPC décroît légèrement, l'ILS_GFK croît de plus de 4 points, ce qui est assez surprenant. En effet la meilleure prise en compte des substitutions à travers la mise à jour plus rapide des pondérations devrait a priori déboucher sur une diminution d'indice, lors du passage du chaînage annuel au chaînage bimestriel. Si l'on observe le phénomène inverse, c'est certainement dû à la différence de population (en termes de séries prises en compte dans le calcul) entre les deux mesures. Le chaînage annuel élimine en effet environ 40% des séries pour l'ILS_GFK, et 55% pour l'ILS_IPC. Ces séries sont celles de courte durée de vie, et on a vu précédemment que les séries de cette nature tiraient vers le haut l'indice GFK. Apparemment cet effet de sélection, lorsqu'on procède au chaînage annuel, est moindre pour l'ILS_IPC que pour l'ILS_GFK.

Considérer le chaînage bimestriel permet de prendre en compte la quasi-totalité des séries pour les deux indices. L'écart entre les deux indices ainsi chaînés se réduit notablement (de 5,2 à 1,6 points) lorsqu'on ne pondère plus les séries. Les pondérations de l'ILS_GFK sont issues de chiffres d'affaires réels, donc par définition à jour. Celles de l'ILS_IPC sont déterminées de façon empirique et évoluent en cours d'année au gré des remplacements opérés par les enquêteurs, à un rythme sans doute moins rapide que celles des données GFK. Ainsi pourrait s'expliquer une bonne partie des 3,6 points d'écart dûs à la prise en compte des pondérations.

L'écart résiduel de 1,6 points s'apparente, en première analyse, à un effet prix «pur». Quelques comparaisons menées sur la distribution des indices de séries sur une même période ont d'ailleurs confirmé que celle de l'ILS_GFK était décalée vers le bas par rapport à celle de l'ILS_IPC. Cependant une partie de cet écart a probablement une origine différente, liée à la moindre substitution entre séries dans l'indice IPC par rapport à l'indice GFK (voir partie 2.2.4.). Autrement dit le véritable effet prix «pur» est vraisemblablement modeste, inférieur à 1,6 point, mais il conviendra bien sûr de vérifier ce résultat par des analyses ultérieures, plus directes ou prenant en compte l'ensemble des séries avec un traitement des effets qualité.

CONCLUSION

Notre étude fournit un premier bilan de ce que représente pour un office national de statistiques l'utilisation de micro-données. Un premier constat est le fort coût d'entrée, du fait de plusieurs phases dont on ne peut s'abstraire : une familiarisation avec l'univers des produits, d'autant plus ardue que ces derniers sont « techniques » ; le traitement d'un volume important de données ; le recours à des modes automatiques ou manuels de détection puis de correction d'erreurs. Ces phases constituent un obstacle majeur à une éventuelle mise en production des micro-données. Elles s'ajoutent à une périodicité et un délai de mise à disposition encore trop longs par rapport à ce qu'exige l'IPC. De plus, le niveau d'agrégation des données, rendu obligatoire pour préserver la confidentialité, occulte les notions de point de vente et d'agglomération. Ainsi, si l'on envisageait de calculer mensuellement l'IPC à partir de ces micro-données, il faudrait disposer de données de meilleure qualité, plus rapidement disponibles et mieux localisées, sauf à envisager un changement de méthode sur le niveau de calcul et d'agrégation des micro-indices.

Il faut reconnaître cependant que ce type de données a l'avantage de fournir de l'information relative aux quantités consommées par modèle (information que l'IPC ne possède pas), et qu'il représente une estimation de l'exhaustivité du marché. En ce sens les micro-données peuvent aider à l'amélioration de la qualité de notre indice de prix. D'ailleurs le cas des téléviseurs et des lave-vaisselle est concluant : en se servant des données GFK comme d'une « référence », nous avons pu évaluer la représentativité de l'échantillon IPC.

Ainsi, il apparaît que le niveau des prix collectés, les répartitions par forme de vente et par marque sont assez satisfaisants, mais que la mise à jour de l'échantillon tarde à prendre en compte la rapidité des évolutions du marché. Nous avons également mis en évidence des défauts de couverture sur des classes de produits (les lave-vaisselle intégrables, les téléviseurs 16/9^{ème} par exemple). Enfin nous nous sommes rendus compte que la sélection des modèles par les enquêteurs suivait essentiellement la règle du « bien suivi, bien vendu ». Les défauts de couverture mis au jour doivent bien sûr être revus. L'obtention d'échantillons de téléviseurs et de lave-vaisselle plus adaptés à la réalité du marché sera facilitée par la gestion centralisée des remplacements dans le secteur Biens Durables. Cependant, il n'est pas pertinent d'équilibrer ces échantillons sur la totalité des critères concernés. En particulier, une distribution uniforme des modèles ne serait pas souhaitable : la contrainte de suivi empêche la collecte des prix de modèles trop instables. Les premiers calculs d'indices montrent des différences entre les deux sources de données, avec des baisses plus prononcées pour les données GFK. Cette situation résulte en partie d'effets prix et d'effets substitution, cependant l'absence de prise en compte des effets qualité restreint le cadre d'analyse, et incite à approfondir l'étude.

Nous n'avons abordé ici qu'une infime part des explorations théoriques rendues possibles par les données scannées. Ces dernières peuvent encore nous apporter beaucoup, notamment dans le domaine des sondages, des ajustements de qualité et du comportement économique des consommateurs.

ANNEXE : DICTIONNAIRE DES DONNEES

N°1 : Liste des variables GFK

TELEVISEURS

VARIABLE	TYPE	MODALITES
<i>marque</i>	caractère	
<i>code GFK</i>	caractère	
<i>mar</i>	code numérique	
<i>numordre</i>	code numérique	
<i>an_intro</i>	quantitatif	de 1994 à 2001
<i>cartepc</i>	qualitatif	'0' sans carte pc '1' avec carte pc sans modem '2' avec carte pc et modem '3' sans carte pc / avec modem internet
<i>dolby</i>	qualitatif	'0' sans dolby '1' avec dolby prologic / surround '2' avec dolby digital '3' avec virtual dolby
<i>hertz</i>	qualitatif	'0' sans option 100 hertz '1' avec option 100 hertz
<i>combi</i>	qualitatif	'0' TV seule '1' avec CD-I '2' avec CD-vidéo '3' avec DVD
<i>nicam</i>	qualitatif	'0' sans '1' avec
<i>recept</i>	qualitatif	'1' pal '2' pal secam '3' secam '4' PSNTSC '5' PSD2MAC '6' MAC '7' pal plus '8' pal plus secam '9' pal plus NTSC
<i>satellit</i>	qualitatif	'0' sans '1' avec
<i>svhs</i>	qualitatif	'0' sans '1' avec
<i>télétexte</i>	qualitatif	'0' sans '1' avec
<i>tuner</i>	qualitatif	'0' sans '1' satellit tuner analogique
<i>pouce</i>	quantitatif	de 4 à 52
<i>type</i>	qualitatif	'0' table télécommande stéréo '1' portable télécommande stéréo enceintes incorporées '2' portable mono '3' table télécommande mono '4' portable télécommande mono '5' portable avec radio stéréo '6' table télécommande stéréo sans enceintes incorporées
<i>taille</i>	qualitatif	'1' de 0 à 35 cm '2' de 36 à 39 cm '3' de 40 à 44 cm '4' de 45 à 48 cm '5' de 50 à 54 cm '6' de 55 à 56 cm '7' de 61 à 69 cm '8' 70 cm et plus
<i>format</i>	qualitatif	'0' 4/3 '1' 16/9 ou 16/9+
<i>tube</i>	qualitatif	'1' normal '2' FST '3' RFT
<i>fv</i>	qualitatif	'1' total des formes de vente '2' hypermarchés '3' grands et multispécialistes '4' magasins traditionnels '5' grands magasins et VPC

LAVE-VAISSELLE

VARIABLE	TYPE	MODALITES
<i>marque</i>	caractère	
<i>code GFK</i>	caractère	
<i>mar</i>	code numérique	
<i>numordre</i>	code numérique	
<i>an_intro</i>	quantitatif	de 1989 à 1999
<i>an_nais</i>	quantitatif	de 1985 à 2001
<i>mois_nais</i>	quantitatif	de 1 à 12
<i>hauteur</i>	quantitatif	de 20 à 90 cm
<i>largeur</i>	quantitatif	de 40 à 87 cm
<i>couvert</i>	quantitatif	de 4 à 14 couverts
<i>nbprog</i>	qualitatif	'0' 1 programme '1' 2 programmes '2' 3 programmes '3' 4 programmes '4' 5 programmes '5' 6 programmes '6' plus de 6 programmes
<i>pan_reg</i>	qualitatif	'0' avec panier réglable '1' sans panier réglable
<i>litre</i>	quantitatif	de 7 à 180 litres d'eau consommés
<i>kwh</i>	quantitatif	de 0 à 3 kwh consommés
<i>decibel</i>	quantitatif	de 43 à 63 decibels
<i>nb_tempe</i>	qualitatif	'0' 1 niveau '1' 2 niveaux '2' 3 niveaux '3' 4 niveaux '4' 5 niveaux
<i>type</i>	qualitatif	'0' lave-vaisselle seul '1' lave-vaisselle + seconde fonction '2' lave-vaisselle + 2 autres fonctions
<i>toucheco</i>	qualitatif	'1' avec '2' sans
<i>dep_dif</i>	qualitatif	'0' sans option départ différé '1' avec option départ différé
<i>cl_power</i>	qualitatif	'1' classe d'efficacité énergétique A '2' classe d'efficacité énergétique B '3' classe d'efficacité énergétique C '4' classe d'efficacité énergétique D '5' classe d'efficacité énergétique E '6' classe d'efficacité énergétique F '7' classe d'efficacité énergétique G '9' classe d'efficacité énergétique indéfinie
<i>cl_lav</i>	qualitatif	'1' classe d'efficacité de lavage A '2' classe d'efficacité de lavage B '3' classe d'efficacité de lavage C '4' classe d'efficacité de lavage D '5' classe d'efficacité de lavage E '6' classe d'efficacité de lavage F '7' classe d'efficacité de lavage G '9' classe d'efficacité de lavage indéfinie
<i>cl_sech</i>	qualitatif	'1' classe d'efficacité de séchage A '2' classe d'efficacité de séchage B '3' classe d'efficacité de séchage C '4' classe d'efficacité de séchage D '5' classe d'efficacité de séchage E '6' classe d'efficacité de séchage F '7' classe d'efficacité de séchage G '9' classe d'efficacité de séchage indéfinie
<i>fv</i>	qualitatif	'1' total des formes de vente '2' hypermarchés '3' grands et multispécialistes '4' magasins traditionnels '5' grands magasins et VPC

N°2 : Liste des variables retenues

TELEVISEURS

VARIABLE	TYPE	MODALITES
<i>marque</i>	caractère	
<i>référence</i>	caractère	
<i>an_intro</i>	quantitatif	de 1994 à 2001
<i>cartepc</i>	qualitatif	'0' sans carte pc '1' avec carte pc sans modem '2' avec carte pc et modem '3' sans carte pc / avec modem internet
<i>dolby</i>	qualitatif	'0' sans dolby '1' avec dolby prologic / surround '2' avec dolby digital '3' avec virtual dolby
<i>hertz</i>	qualitatif	'0' sans option 100 hertz '1' avec option 100 hertz
<i>combi</i>	qualitatif	'0' TV seule '1' avec CD-I '2' avec CD-vidéo '3' avec DVD
<i>nicam</i>	qualitatif	'0' sans '1' avec
<i>recept</i>	qualitatif	'1' pal '2' pal secam '3' secam '4' PSNTSC '5' PSD2MAC '6' MAC '7' pal plus '8' pal plus secam '9' pal plus NTSC
<i>satellit</i>	qualitatif	'0' sans '1' avec
<i>svhs</i>	qualitatif	'0' sans '1' avec
<i>télétexte</i>	qualitatif	'0' sans '1' avec
<i>tuner</i>	qualitatif	'0' sans '1' satellit tuner analogique
<i>DIAG</i>	quantitatif	de 11 à 132 cm
<i>TYP</i>	qualitatif	'P M' portable mono 'PCM' portable télécommande mono 'PCS' portable télécommande stéréo 'TCM' table télécommande mono 'TCS' table télécommande stéréo
<i>FOR</i>	qualitatif	'4/3' 4/3 '16/9' 16/9 ou 16/9+
<i>TUB</i>	qualitatif	'NOR' normal 'FST' FST 'RFT' RFT
<i>fv</i>	qualitatif	'1' total des formes de vente '2' hypermarchés '3' grands et multispécialistes '4' magasins traditionnels '5' grands magasins et VPC

LAVE-VAISSELLE

VARIABLE	TYPE	MODALITES
<i>marque</i>	caractère	
<i>référence</i>	caractère	
<i>code</i>	caractère	
<i>POS</i>	qualitatif	'FS' pose libre 'BI' encastrable/habillable 'IE' intégrable
<i>an_naiss</i>	quantitatif	de 1985 à 2001
<i>hauteur</i>	quantitatif	de 20 à 89 cm
<i>largeur</i>	quantitatif	de 43 à 60 cm
<i>couvert</i>	quantitatif	de 4 à 14 couverts
<i>nbprog</i>	qualitatif	'0' 1 programme '1' 2 programmes '2' 3 programmes '3' 4 programmes '4' 5 programmes '5' 6 programmes '6' plus de 6 programmes
<i>pan_reg</i>	qualitatif	'0' avec panier réglable '1' sans panier réglable
<i>l1</i>	quantitatif	de 7 à 50 litres d'eau consommés
<i>l2</i>	quantitatif	de 12 à 22 litres d'eau consommés
<i>k1</i>	quantitatif	de 0.6 à 2.4 kwh consommés
<i>k2</i>	quantitatif	de 0.7 à 1.7 kwh consommés
<i>decibel</i>	quantitatif	de 43 à 63 decibels
<i>nb_tempe</i>	qualitatif	'0' 1 niveau '1' 2 niveaux '2' 3 niveaux '3' 4 niveaux '4' 5 niveaux
<i>type</i>	qualitatif	'0' lave-vaisselle seul '1' lave-vaisselle + seconde fonction '2' lave-vaisselle + 2 autres fonctions
<i>toucheco</i>	qualitatif	'1' avec '2' sans
<i>dep_dif</i>	qualitatif	'0' sans option départ différé '1' avec option départ différé
<i>cl_power</i>	qualitatif	'1' classe d'efficacité énergétique A '2' classe d'efficacité énergétique B '3' classe d'efficacité énergétique C '4' classe d'efficacité énergétique D '5' classe d'efficacité énergétique E '6' classe d'efficacité énergétique F '7' classe d'efficacité énergétique G '9' classe d'efficacité énergétique indéfinie
<i>cl_lav</i>	qualitatif	'1' classe d'efficacité de lavage A '2' classe d'efficacité de lavage B '3' classe d'efficacité de lavage C '4' classe d'efficacité de lavage D '5' classe d'efficacité de lavage E '6' classe d'efficacité de lavage F '7' classe d'efficacité de lavage G '9' classe d'efficacité de lavage indéfinie
<i>cl_sech</i>	qualitatif	'1' classe d'efficacité de séchage A '2' classe d'efficacité de séchage B '3' classe d'efficacité de séchage C '4' classe d'efficacité de séchage D '5' classe d'efficacité de séchage E '6' classe d'efficacité de séchage F '7' classe d'efficacité de séchage G '9' classe d'efficacité de séchage indéfinie
<i>fv</i>	qualitatif	'1' total des formes de vente '2' hypermarchés '3' grands et multispecialistes '4' magasins traditionnels '5' grands magasins et VPC

BIBLIOGRAPHIE

J. Bascher and T. Lacroix (1998), « De l'utilisation des méthodes hédoniques dans l'IPC : application aux biens durables et à l'habillement », *Contribution à la 4^{ème} conférence internationale du groupe d'Ottawa, Washington, Avril 1998.*

J. Bascher and T. Lacroix (1999), « Lave-vaisselle et micro-ordinateurs dans l'IPC français : lamodélisation hédonique, de la théorie à la pratique », *Contribution à la 5^{ème} conférence internationale du groupe d'Ottawa, Reykjavik, Août 1999.*

Catalogues constructeurs (les principaux utilisés)

Pour les lave-vaisselle : Brandt, Bosch, Whirlpool, Siemens

Pour les télévisions : Thomson, Philips, Sony, Grundig

D.Fenwick and A. Ball (2001), « Sampling in Consumer Price Indices : what role for scanner data ? », *Sixth meeting of the International Working Group on Price Indices, Australia 2001.*

GFK (1997), « Le suivi quantitatif des marchés des biens d'équipement et de loisirs audiovisuels, note méthodologique ».

INSEE (1998), « Pour comprendre l'Indice des prix », *INSEE méthodes n°81-82.*

R. Lowe (1998), « Televisions : quality changes and scanner data », *Paper for the Ottawa Group, Washington, April 1998.*

R. Lowe (2001), « Téléviseurs : variations de qualité et données scannographiques », *Statistique Canada, Série analytique n°14.*

F. Magnien and J.Pougnard (1998), « Etude du chaînage d'indices de prix à l'aide de micro-données », *INSEE Méthodes, n°84-85-86.*

H. Scobie (1998), « Utilisations possibles des données scannographiques ; Etude de cas à l'aide des données sur le café », *Statistique Canada, Série analytique n°6.*

P. Tassi (1989), « Méthodes Statistiques », *Economica.*