Le projet d'utilisation des données de caisse de la grande distribution alimentaire pour le calcul de l'Indice des Prix à la Consommation français

Sébastien Faivre (*)

(*) Insee, Division des Prix à la Consommation

Les progrès techniques permettent d'enregistrer de façon exhaustive et précise l'ensemble des articles passant à la caisse des magasins de la grande distribution. Ces informations sont bien sûr utilisées par les magasins eux-mêmes (gestion de stock en temps réel, suivi du marché…) mais sont également mobilisées par des entreprises spécialisées dans le marketing et les études de marché comme Nielsen.

Ces sources sont pour l'instant peu utilisées par l'Insee. La comptabilité nationale et la division IPC ont recours à des données agrégées afin de suivre l'évolution des quantités consommées et les parts de marché des circuits de distribution. Par contre aucune donnée élémentaire n'est exploitée.

Le projet vise à utiliser le potentiel important des données élémentaires mobilisables, afin de fournir des données de base pour le calcul de l'indice des prix permettant d'améliorer sa précision (données exhaustives sur le champ couvert), tout en modernisant les processus de collecte et en renforçant leur fiabilité.

D'autre part, les données de caisse, compte tenu de leur richesse et de leur caractère exhaustif, constituent une source d'information de grande valeur pour enrichir la connaissance des prix et des comportement des consommations : en plus du calcul de l'indice des prix, elles pourraient ainsi être utilisées pour mettre en place des indices de prix moyen (préconisation du rapport du CAE sur la mesure du pouvoir d'achat), faire des comparaisons spatiales de prix sur le champ géographique couvert (métropole hors Corse), ou encore permettre un suivi fin de la consommation des ménages complétant l'enquête Budget des Familles.

Il s'agit cependant d'une source de données nouvelle qui nécessite une phase d'exploration approfondie avant de déterminer s'il est possible de l'utiliser ou non pour le calcul de l'indice des prix.

De ce fait, une étude méthodologique importante a été menée de septembre 2009 à mars 2011 pour déterminer la faisabilité de l'utilisation de ces données.

La première partie présente brièvement la source données de caisses. Le seconde partie détaille les premiers travaux de simulations d'indices de prix réalisés à partir des données de caisse dans le cadre de l'étude de faisabilité. Il s'agit encore à ce stade de premiers résultats provisoires, qui seront affinés et complétés au cours de l'année 2012.

Première partie : Le cadre méthodologique de l'utilisation des données de caisse dans l'indice des prix

1. Présentation de la source de données

Les données de caisses (« scanner data ») sont directement issues des données détaillées enregistrées aux caisses sur les achats effectués par les clients de la grande distribution.

On connaît alors pour chaque transaction élémentaire la nature du produit acheté (repéré sauf cas particulier par son code-barres sur 13 positions), le nombre d'unités vendues et le prix total payé par l'acheteur.

Pour chaque référence vendue dans le magasin (repérée par son code-barres), on dispose alors dans les enregistrements de caisses du nombre d'unités vendues et du prix total payé pour ces unités vendues au niveau journalier[1], le triplet (code-barres, nombre d'unités vendues, ventes totales en valeur) formant la donnée élémentaire reçue des points de ventes de la grande distribution.

Remarque : Afin de limiter les volumes de données, les ventes et les quantités vendues sont généralement agrégées à la semaine par les deux entreprises collectant ces données : on dispose alors pour chaque code-barres d'un prix de vente hebdomadaire moyen.

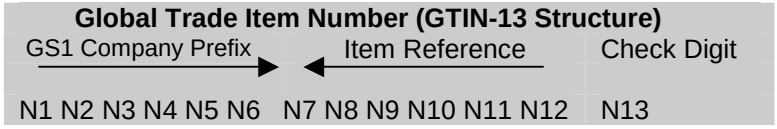
Le code-barres[2]

L'identification, lors de leur passage en caisse, des produits vendus dans les grandes enseignes de la distribution alimentaire s'effectue à l'aide des codes à barres qui figurent sur chaque article. A chaque code-barres correspond une séquence de 13 chiffres formant l'EAN (European Article Numbering)[3] qui constitue l'identifiant individuel des données de caisse.

En France ce système de numérotation des articles est piloté par la société GS1 France, membre de l'association internationale GS1 spécialisée notamment dans la conception et l'harmonisation des conditions de mise en œuvre de standards d'identifiants des produits.

Toute entreprise qui adhère en France au système d'identifiants par code barres -EAN se voit attribuer par GS1 France, une plage d'EAN qu'elle peut utiliser librement pour identifier ses produits, dans le respect des règles d'usage prescrites par GS1.

La composition du numéro est décrit dans le schéma ci dessous proposé par GS1.



Les trois premiers chiffres de l'EAN indiquent *le pays* dans lequel l'entreprise a adhéré au système EAN. Par exemple les trois premiers chiffres des numéros attribués par l'intermédiaire de GS1-France sont compris entre 300 et 379 (ceux attribués par GS1 Germany entre 400 et 440 et par GS1 China entre 690 à 695).

Le code formé par les six premiers chiffres -ou préfixe du code barres constitue le numéro mondial d'identification de *l'entreprise* adhérente de GS1. Ce code barre permet ainsi l'identification individuelle du fabricant du produit correspondant.

[1] Le prix associé à un code-barres restant toujours constant au cours d'une même journée.

[2] Cf. rapport du groupe de travail Enseignes/Insee, page 14

[3] Dénommé GTIN (Global Trade Item Number) par GS1

Le code formé par les douze premiers chiffres est un identifiant du *produit*. Le treizième chiffre est une clé de contrôle.

Ces numéros sont utilisés comme identifiants des produits dans les catalogues des fabricants, dans les transactions entre fabricants et distributeurs, dans les référentiels de produits des enseignes, et dans les bases centralisant les données de caisse : selon des éléments fournis par la société IRI, plus de 1600 nouveaux codes-barres sont créés chaque semaine[4]

Il convient de noter que les produits frais à poids variable et, de façon générale, les produits ou articles qui donnent lieu à conditionnement dans le point de vente - c'est-à-dire ceux qui ne relèvent pas, pour les distributeurs, d'une activité de pur négoce - sont étiquetés de "faux EAN ", qui en permettent l'identification lors du passage en caisse, au moyen de référentiels et des classifications propres au point de vente ou à l'enseigne.

Les données de caisse présentent de manière évidente des avantages importants par rapport à la collecte prix traditionnelle, dans la mesure où ils permettent **d'observer un échantillon de prix dans la grande distribution beaucoup plus large que dans la collecte enquêteurs[5], de s'affranchir des contraintes liées à la collecte** (observation des prix dans un nombre restreint d'agglomérations…) **et d'éviter les erreurs éventuelles dans la saisie des prix par les enquêteurs**.

En pratique, une difficulté pour utiliser le code-barres comme identifiant statistique du produit vient cependant du fait qu'il **comprend un identifiant produit attribué par le fabricant selon des règles de gestion qui lui sont propres** (même si ces règles sont encadrées par un ensemble de « bonnes pratiques » développées par la société GS1, responsable la standardisation des échanges entre les acteurs de la grande distribution).

On notera également que certains produits vendus en supermarchés ne disposent pas de codes-barres standardisés (fruits et légumes, produits à la coupe, viandes emballées sur place, baguettes ou pâtisseries confectionnées sur place)[6]. Ces produits seraient donc susceptibles de continuer à être suivis par des enquêteurs dans les supermarchés.

2. Problématiques méthodologiques liées à l'utilisation des données de caisse

a. Choix des indices utilisés sur le champ couvert par les données de caisse

Selon la théorie économique, l'indice des prix correspond à la variation du budget minimal nécessaire pour que le consommateur puisse maintenir son niveau d'utilité entre la période de référence et la période courante. Il n'existe donc pas un indice de prix unique, mais différentes formes d'indices de prix, en fonction des hypothèses effectuées sur la fonction d'utilité et du choix de la période de référence pour la pondération des variations de prix élémentaires. La question de la fréquence du chaînage (mise à jour du panier de biens suivis et des pondérations) est également essentielle.

L'indice des prix français actuel est un indice de Laspeyres, compte tenu du fait qu'on ne dispose pas d'informations sur les ventes de la période courante. Cet indice tend par construction à surestimer - légèrement - les hausses de prix.

L'utilisation des données de caisse permet cependant de lever cette contrainte d'information absente sur la période courante, puisqu'on dispose d'informations sur les ventes à chaque période. Il est alors possible de calculer sur le champ couvert par les données de caisse des indices superlatifs de type

[4] Hors le cas très particulier des vins AOC, pour lesquels 1000 nouveaux codes-barres sont créés chaque semaine.

[5] Par exemple, la Norvège suit désormais 14 000 références dans le secteur de l'alimentaire et des boissons non alcoolisées grâce aux données de caisse alors que seules 250 références étaient suivies dans la collecte traditionnelle.

[6] Le code-barres qui apparaît sur l'emballage des produits est alors un code local (codes barres librement utilisables dont les trois premiers chiffres sont compris entre 020 et 029) qui n'a de sens qu'à l'intérieur du magasin.

Fisher ou Tornqvist, qui ne présentent pas de biais systématique contrairement aux indices de Laspeyres et de Paasche.

Cependant, on ne dispose pas d'un recul suffisant sur ces nouveaux indices pour envisager de les utiliser à la place de l'indice de Laspeyres. Par ailleurs, un tel choix serait contraire au règlement cadre européen[7], qui précise à l'article 9 que « les États Membres traitent les données collectées afin de produire l'IPCH sur la base d'un indice de Laspeyres ».

b. Fréquence du chaînage sur le champ couvert par les données de caisse

Le chaînage de l'indice correspond à la mise à jour du panier de biens suivis et des pondérations des indices élémentaires.

Actuellement, le chaînage de l'indice français est effectué à un rythme annuel, au mois de décembre. Lorsqu'un produit suivi disparaît entre deux opérations de chaînage, il est remplacé par un produit de remplacement le plus « proche possible » du produit disparu (en termes de caractéristiques du produit et du point de vente).

Cependant, les données scannées s'adaptent plus difficilement à un chaînage annuel compte tenu de la très forte attrition des codes-barres : ainsi, sur l'ensemble du champ de l'alimentaire, on observe sur des données de caisse norvégiennes de 2004 que 27% des codes-barres ont disparu au bout d'un mois et 58% au bout d'un an[8].

Ainsi, en cas de maintien du mode de calcul actuel « traditionnel » de l'indices avec chaînage annuel et utilisation d'indices de type Laspeyres, le nombre de replacements de produits à effectuer s'accroîtra très fortement par rapport à la situation actuelle, puisqu'il pourrait concerner la moitié des produits. Des remplacements de produits à une échelle aussi grande sont susceptibles de représenter une forte source d'instabilité pour les indices de prix, et une charge de gestion importante pour identifier les produits de remplacement. Par ailleurs, l'approche par produits de remplacement nécessiterait une documentation très précise des codes-barres en termes de marque, de caractéristiques techniques….

Ainsi, une approche alternative « novatrice » de mise à jour en continu du panier de référence par chaînage mensuel a été développée. Cependant, sur le plan théorique, un chaînage trop fréquent entraîne des phénomènes de dérives des indices[9].

En règle générale, dans le cas d'indices « superlatifs » (c'est-à-dire proches d'un indice à utilité constante et ne présentant de tendance systématique à la surestimation ou à la sous-estimation de l'évolution des prix), comme le Fisher ou le Tornqvist, on constate plutôt un effet de dérive à la baisse des indices (de Haan et van der Grient, 2009), particulièrement évidente dans le cas d'un chaînage hebdomadaire. Les résultats observés pour un chaînage mensuel semblent beaucoup plus plausibles, mais l'absence de dérive reste néanmoins à démontrer.

Pour des raisons de lisibilité et de crédibilité de l'indice, l'option retenue par l'Insee est de reconduire la méthode actuelle de panier fixe avec chaînage annuel sur les données de caisse.

Un tel choix assure que les indices produits à partir des données de caisse seront conformes à la réglementation internationale et européenne en matière d'élaboration d'indices de prix[10].

[7] Règlement (CE) 2494/95 du 23 octobre 1995.

[8] Ce phénomène est confirmé dans d'autre pays. Ainsi, sur des données de caisse hollandaises de ventes de papier toilette sur la période 2005-2008, on constate qu'environ la moitié des articles suivis ont disparu au bout d'un an.

[9] Le phénomène de dérive des indices se produit lorsque alors que l'indice de prix ne retrouve pas sa valeur initiale, après une hausse ou une baisse temporaire (phénomène de soldes par exemple), alors que le vecteur des prix de l'économie retourne à sa position initiale. Un tel phénomène est susceptible d'entraîner des biais importants dans la mesure de l'indice des prix.

[10] La seule nuance par rapport aux indices issus de la collecte enquêteurs vient du fait qu'en cas de remplacement du produit, on dispose dans les données de caisse du prix au mois de base du produit remplaçant, ce qui permet d'éviter une comparaison directe entre le prix du produit remplaçant et celui du produit remplacé.

Deuxième partie : simulations d'indices de prix à partir des données de caisses

Les résultats présentés ici constituent des premiers résultats provisoires. Ils seront complétés et affinés dans le cadre des travaux menés par le CPS adjoint au cours de l'année 2012.

La collecte « traditionnelle » de l'indice des prix à la consommation se base sur le suivi mensuel du prix d'un panier de biens et services. Dans ce cadre, un produit est défini par le croisement entre un « article » (par exemple le camembert de 250g de marque A) et un point de vente (l'hypermarché B). Un produit appartenant au panier de référence suivi chaque mois pourra être alors par exemple le camembert de 250g de marque A dans le point de vente B.

La prise en compte du point de vente dans la définition du produit vient du fait qu'on considère que les points de ventes ne rendent pas le même service au consommateur. En particulier, les commerces de proximité ne rendent pas le même service au consommateur que les grands hypermarchés, moins chers mais d'un temps d'accès beaucoup plus long.

L'objectif prioritaire de l'étude de faisabilité est de transposer la méthode actuelle de suivi d'un panier fixe chaîné annuellement dans le cadre des données de caisses.

Dans les données de caisses, les produits sont identifiés par le code-barres. Il s'agit donc d'un concept un peu différent, puisqu'un même produit peut en théorie être présent sous différents codes-barres[11] dans certains cas (par exemple si un prix conseillé est affiché sur l'emballage dans certains points de ventes et pas dans d'autres).

De manière analogue au « produit » défini dans le cadre de la collecte traditionnelle, on définit pour les données de caisses la notion de **« série »** comme le **croisement entre un code-barres et un point de vente**. Un exemple de série sera alors le camembert de 250g immatriculé sous le code-barres C dans le point de vente B.

Cette note étudie le calcul d'indices de prix à partir d'un panier de codes-barres, à partir de l'échantillon de données de test livré par la société IRI début novembre 2010 et portant sur les données de ventes hebdomadaires pour :

- 10 grandes familles de produits (déclinées en 17 familles élémentaires IRI)
- 1000 points de ventes des enseignes participant au test[12]
- 3 années : 2007 à 2009.

1. Peut-on suivre dans le panier de référence l'ensemble des codes-barres ? La question de la stabilité des codes-barres

Une difficulté importante pour la constitution et le suivi d'un panier de codes-barres vient de l'instabilité des codes-barres. Selon une étude menée sur données norvégiennes en 2004 sur le champ de l'alimentaire, 27% des codes-barres ont disparu au bout d'un mois et 58% au bout d'un an.

Afin de tester ce phénomène sur données françaises, une première étude sur la stabilité des codes-barres a été effectuée à partir de l'échantillon de test sur l'année 2009. Pour ce faire, on a inclus dans

[11] Le test terrain « données de caisses » montre cependant que ce phénomène est très limité ou inexistant, puisque pour les 259 produits bien appariés on avait soit un seul codes-barres, soit plusieurs codes-barres mais qui différaient sur au moins une caractéristique. Plus généralement, aucun cas de codes-barres différents mais présentant exactement les mêmes caractéristiques n'a été détecté lors du test.

[12] Enseignes représentant environ 30% de la surface de vente totale de la grande distribution alimentaire.

l'échantillon annuel 2009 l'ensemble des séries présentes au mois de base (mois de décembre 2008), soit en tout 1 096 000 séries pour les 17 familles de produits IRI suivies.

On a ensuite regardé parmi tous les codes-barres présents en décembre 2008 ceux qui étaient stables au cours de l'année 2009, c'est-à-dire ceux dont les ventes étaient strictement positives pour chacun des mois de l'année 2009.

On obtient par famille de produits IRI les résultats suivants :

famille IRI	Description famille IRI	Nb Séries (croisement code-barres et point de vente) au mois de base décembre 2008	Ventes totales 2009 des séries présentes au mois de base (en milliers d'euros)	Taux de stabilité des séries sur l'année 2009
205	Cafe en dosette	74736	175 155	39,2%
701	Cafe moulu avec caféine	94518	418 970	53,5%
703	Cafe moulu sans caféine	12343	30 405	54,4%
2628	Papier toilette	24882	295 741	57,3%
3202	Huile cuisine/salade	35048	186 140	68,1%
3203	Huile olive	29505	159 340	45,2%
4206	Riz nature	72831	154 562	58,0%
5112	Pizzas surgelées	29955	120 139	40,7%
5118	Quiches surgelées	17209	29 934	46,3%
5701	Yaourt	217793	793 359	59,3%
6402	Œufs	23792	400 937	70,7%
7904	Chocolat en tablette	198473	455 958	50,7%
8002	Jus de fruits	148115	559 252	54,6%
8271	Fromage pâte persillée	19245	86 802	61,9%
8272	Autre fromage pâte molle	6673	19 315	50,1%
8273	Fromage pâte molle croûte lavée	22047	75 750	57,5%
8274	Fromage pâte molle croûte fleurie	69439	458 457	69,0%
Ensemble		**1 096 604**	**4 420 221**	**55,1%**

On observe ainsi sur les 17 familles de produits IRI un taux de stabilité globale annuelle de 55%, supérieur de 13 points à celui obtenu sur données norvégiennes. Le taux de stabilité des codes-barres est très variable selon les familles, puisqu'il varie de 40% pour les familles les plus volatiles (café en dosettes et pizzas surgelées) à 70% pour les familles les plus stables (huiles de cuisine, œufs, fromages à pâte molle croûte fleurie)[13].

Il ne semble donc pas possible de sélectionner l'ensemble des séries dans l'échantillon annuel, compte-tenu de l'importance des remplacements à effectuer, qui concerneraient 45% des séries. Ainsi, sur la base d'un échantillon global de 1000 points de ventes, d'un nombre moyen de 10 000 codes-barres par magasin et d'un taux de remplacement des codes-barres de 45%, le nombre de

[13] Toutes familles confondues, les résultats obtenus sur l'année 2008 sont relativement similaires à ceux de 2009 : part des codes-barres stables de 55,8% en 2008 contre 55,1% en 2009. Au niveau « famille IRI », les fluctuations sont plus importantes (avec des écarts limités pour les « grandes familles » de produits et des variations plus importantes pour les petites familles), ces variations tendant cependant à se compenser au final.

remplacements à effectuer serait 4 500 000 chaque année, ce qui dépasse visiblement les moyens humains qu'il est possible d'y affecter même en cas d'automatisation importante des traitements.

En revanche, les codes-barres stables correspondent généralement à des codes-barres bien vendus : ainsi, la part dans les ventes des codes-barres stables est de l'ordre de 70%[14] alors que la part en nombre des codes-barres stables n'est que de 55%. Dans ces conditions, une stratégie de sélection des codes-barres fondée sur un tirage proportionnel aux ventes (en application du principe « bien suivi, bien vendu » qui prévaut pour le choix des produits actuellement suivis dans la collecte enquêteurs) semble susceptible de limiter les remplacements à effectuer.

2. Constitution d'un panier de codes-barres par sélection d'un échantillon de séries proportionnellement aux ventes

On teste ici empiriquement au moyen de simulations la constitution d'un panier représentatif de codes-barres

2.1 La taille du panier de codes-barres

L'un des objectifs de l'utilisation des données de caisses est d'augmenter fortement la taille du panier suivi. Dans la collecte IPC actuelle, 50 000 prix sont relevés chaque mois dans la grande distribution (y compris magasins populaires). En considérant que les 10 familles suivies représentent de l'ordre de 10%[15] des ventes de la grande distribution et que les enseignes suivies représentent 30% des ventes de la grandes distribution, le nombre de relevés « enquêteurs » théoriques effectués actuellement pour les 10 familles de produits suivis dans les enseignes participant au test représente 10% x 30% = 3% des relevés mensuels en grande distribution, soit donc environ 1500 relevés.

Ce nombre est à rapporter au nombre d'1,1 millions de séries présentes dans les 1000 points de ventes couverts par le test (soit donc un nombre total de l'ordre de 1,5 millions de séries dans les 1872 points de ventes des enseignes participant au test, compte-tenu du fait que les hypermarchés sont couverts exhaustivement dans l'échantillon tandis que les supermarchés font l'objet d'un sondage). ***Pour le champ des 10 familles de produits suivies et des enseignes participant au test, le taux de sondage actuel de l'IPC est donc voisin de 0,1% des séries.***

Cependant, afin de tenir compte des fluctuations possibles sur le terrain de la répartition des relevés par enseigne, on prend ici par prudence un nombre théorique de 2000 relevés effectués actuellement pour les 10 familles de produits suivies dans les enseignes participant au test.

C'est donc par rapport à ce seuil de 2000 relevés dans la collecte actuelle qu'on pourra calibrer l'échantillon de codes-barres.

Une première étude a montré que, sur la base d'une concentration des relevés dans un nombre restreint d'agglomérations de manière à maintenir constant le revenu médian des enquêteurs dans le cadre de la collecte restante, il était nécessaire de doubler le nombre de séries suivis dans le champ des données de caisses pour maintenir constante la précision de l'indice.

On obtiendrait alors un panier de 4000 séries, nombre porté à 5000 pour des raisons de prudence.

Il est cependant possible d'envisager des paniers de taille plus importante, par exemple 10 000 ou 20 000 codes-barres.

Au-delà, le nombre de remplacements à effectuer risque d'être trop important par rapport au gain statistique lié à l'augmentation de la taille de l'échantillon (il est possible néanmoins tester des paniers de taille plus importante, la seule limite étant le temps nécessaire aux simulations).

[14] D'après les simulations de tirages effectuées, cf. infra paragraphe 3

[15] En décembre 2009, les variétés correspondant aux 10 familles de produits suivis représentaient (y compris les ordres de recherche pour l'échantillon 2010) 4 063 produits sur 49 919 relevés, soit 8,1% des relevés.

A titre d'exemple, on étudie ici le tirage d'un échantillon de 10 000 séries, réparties entre les 17 familles IRI proportionnellement au montant total des ventes.

2.2 Le calcul des probabilités d'inclusion

Il est nécessaire de prendre en compte dans le calcul de la probabilité d'inclusion des séries le processus de sélection des 1000 points de ventes de l'échantillon, lié d'une part, au fait qu'IRI ne couvre que 1545 des 1876 points de ventes appartenant aux enseignes de la grandes distribution participant au test, et, d'autre part, au fait qu'on a sélectionné un échantillon de 1000 points de ventes parmi les 1545 points de ventes accessibles.

On calcule donc les probabilités d'inclusion des séries proportionnellement à une variable de ventes redressée, obtenue (dans le cadre d'une stratification enseigne*Hyper/Super) en multipliant le montant des ventes de la série par l'inverse de la part de la surface de vente couverte dans l'échantillon pour la strate considérée.

2.3 Résultats sur l'importance de remplacements à effectuer

2.3.a Stabilité des échantillons de séries observée dans les simulations

On a simulé ici le tirage d'un échantillon de 10000 séries, soit 5 fois la taille du panier minimal, pour l'échantillon annuel 2009.

Les 10000 séries à tirer ont été répartis au sein des 17 famille IRI proportionnellement au volume des ventes en 2009 de chaque famille[16].

On a tiré ensuite dans chaque famille IRI un nombre de séries correspondant à l'allocation calculée à l'étape précédente, avec une probabilité d'inclusion proportionnelle au montant des ventes de la série en décembre 2008 redressé pour prendre en compte le processus de sélection des points de ventes (cf. paragraphe précédent).

Les variables d'équilibrage utilisées sont l'enseigne et la marque[17]. On a essayé d'introduire d'autres variables supplémentaires (par exemple, les variables variété/parfum et quantité), mais cela conduisait à une augmentation trop importante du temps nécessaire pour effectuer les tirages[18].

500 tirages ont été effectués pour la simulation, pour un temps total de simulation de 66 heures[19]. Le tableau ci-dessous donne les résultats observés en termes de stabilité des codes-barres:

Année	Taux de stabilité moyen sur l'ensemble des 10 familles	Min	Q5	Q25	Med	Q75	Q95	Max
2009	**72,3%**	71,1%	71,6%	72,0%	72,3%	72,6%	73,0%	73,5%

On constate ainsi un taux de stabilité des séries sélectionnées de 72% bien supérieur au taux de stabilité moyen des séries de la base de sondage (55%).

Remarque : Le taux de stabilité moyen sur l'année 2008 (70,8%) est très proche mais légèrement inférieur à celui de l'année 2009 (72,3%).

[16] Ventes totales par famille calculées à partir des séries présentes dans la base de sondage 2009.

[17] La notion de marque au sens IRI correspond plutôt à la notion de « référence » au sens de la collecte IPC. Par exemple, pour un Yaourt Danone Taillefine, la marque IRI sera « Taillefine ».

[18] En effet, le temps mis par la macro fastcube pour effectuer les tirages est en $O(p^2 N)$ où p est le nombre de variables d'équilibrage et N la taille de la population. On rappelle que, si on souhaite équilibrer sur la variable marque, il faut introduire dans l'équilibrage autant de variables indicatrices qu'il y a de modalités possibles pour la marque.

[19] Les programmes ayant tourné du vendredi 18h au lundi à 12h.

2.3.b Comparaison avec la collecte IPC

Une étude du taux de remplacements effectués en 2009 dans le cadre de la collecte IPC pour les variétés IPC correspondant aux 10 familles de produits suivies dans les données de test a été effectuée.

On constate ainsi en 2009 **un taux de remplacement de 17%** pour les produits suivis dans le cadre de la collecte **IPC**, à comparer avec un **taux de remplacement de 28%** pour les échantillons annuels de séries issus des **données de caisses**.

2.3.c Une meilleure prise en compte des produits à vie courte dans les échantillons de séries issus des données de caisses

Une première explication de la différence sur les taux de remplacements vient de la prise en compte ou non des promotions « fabricant ». On entend par promotions « fabricant » des produits différents en termes de conditionnement et de packaging du produit habituellement vendu, et dont le prix au litre ou au kg est plus bas que pour le produit standard : par exemple un lot de deux bouteilles avec 15% gratuit au lieu de la bouteille à l'unité vendue habituellement (par opposition aux promotions « magasins » dans lesquelles le magasin baisse le prix du produit habituellement en rayon)

En effet, ces promotions « fabricant » correspondent à des produits (et des codes-barres) temporaires, qui donneront lieu dans la très grande majorité des cas à un remplacement.

En pratique, la prise en compte des promotions « fabricant » [20] constitue une des difficultés de l'IPC[21] : en effet, lorsque le produit en promotion coexiste avec le produit sans promotion, l'enquêteur saisit généralement le prix du produit habituel qui n'est pas en promotion (dans la mesure où il n'est pas possible de saisir deux prix pour un même produit).

En revanche, dans les simulations d'échantillons de codes-barres, les promotions « fabricant » sont traitées comme les autres codes-barres, ce qui fait qu'on retrouve en moyenne dans les échantillons tirés la même proportion de codes-barres en promotions que dans la base de sondage. (Ainsi, le lot de deux bouteilles de soda avec « 15% gratuit » correspond dans les données de caisses à un code-barres différent de la bouteille de soda à l'unité, et donc à une série différente dans la base de sondage)

On a les résultats suivants sur l'importance des codes-barres associés à une promotion « fabricant » dans la base de sondage :

[20] Alors que la prise en compte des promotions « magasin » dans l'IPC ne pose pas de difficultés.

[21] On décrit ici la situation en 2009. Un travail a été engagé depuis par la Division IPC pour améliorer le suivi des promotions.

FAMILLE_IRI	Descriptif famille	Nb total codes-barres différents	Part des codes-barres en promotion	Part des ventes 2009 liées aux promotions
205	Cafe en dosette	464	21,1%	19,2%
701	Cafe moulu avec caféine	1 018	8,7%	25,5%
703	Cafe moulu sans caféine	102	6,9%	16,4%
2628	Papier toilette	279	30,5%	16,4%
3202	Huile cuisine/salade	406	1,5%	0,8%
3203	Huile olive	612	2,1%	2,2%
4206	Riz nature	599	6,2%	2,2%
5112	Pizzas surgelées	342	0,6%	0,1%
5118	Quiches surgelées	189	0,0%	0,0%
5701	Yaourt	1 479	8,0%	8,6%
7904	Chocolat en tablette	1 388	10,6%	10,5%
8002	Jus de fruits	1 730	3,9%	4,0%
8271	Fromage pâte persillée	197	5,1%	7,5%
8272	Autre fromage pâte molle	307	1,3%	0,3%
8273	Fromage pâte molle croûte lavée	499	1,2%	2,6%
8274	Fromage pâte molle croûte fleurie	769	5,6%	6,7%
Ensemble		**10 380**	**7,1%**	**9,6%**

On constate ainsi que les promotions « fabricant » représentent 7% des codes-barres, pour un peu moins de 10% des ventes.

En supposant que toutes les promotions « fabricant » correspondent à des codes-barres instables (durée de vie du produit inférieure à un an), on retrouve des ordres de grandeur cohérent sur les remplacements, puisque le taux de remplacement annuel global dans les données de caisses (28%) est proche de la somme du taux de remplacements dans l'IPC (17%) où les produits suivis excluent dans la très grande majorité des cas les promotions « fabricant » et du taux de produits correspondant à des promotions « fabricant » (7%).

La différence restante (4%) pourrait venir du fait que l'IPC suit uniquement des produits « biens suivis biens vendus », tandis que les échantillons de séries contiennent aussi des codes-barres correspondant à des produits moins vendus et plus instables, même si c'est en faible proportion compte-tenu du montant limité de leurs ventes.

Il ressort donc de cette étude que l'augmentation des remplacements lors du passage de la collecte IPC actuelle à un échantillon de codes-barres reste limitée (taux passant de 17% à 28%) et que cette hausse peut s'expliquer par une meilleure couverture des produits à durée de vie courte, d'une part les promotions « fabricant » et d'autre part des produits plus atypiques avec un faible chiffre d'affaires.

3. Simulations d'indices de prix annuels 2009 à partir des échantillons de séries tirés dans les données de caisses et comparaison avec les indices issus de la collecte IPC

3.1 Méthodologie de calcul d'indices de prix à partir des échantillons de séries

3.1.a La gestion des remplacements des séries

Lorsqu'une série disparaît (c'est-à-dire que le code-barres considéré n'est plus vendu dans le point de vente considéré), il est nécessaire de la remplacer par une autre série.

On définit ici une série stable sur l'année étudiée comme une série pour laquelle les ventes mensuelles sont strictement positives pour chacun des 12 mois de l'année. Dans le cas contraire, la série est instable et il sera nécessaire d'effectuer un remplacement en cours d'année. Une première étude de la base de sondage montre que 45% des séries sont instables, mais que ces séries instables ne représentent que 28% des ventes, et donc 28% des séries dans les échantillons tirés (cf. Première partie de la note).

Il est à noter que la problématique du remplacement se pose différemment dans le cadre des données de caisses et dans la collecte enquêteurs, puisque, dans les données de caisses, on dispose en règle générale du prix du produit remplaçant le mois précédent, ce qui permet d'éviter une comparaison directe entre le prix du produit remplacé et celui du produit remplaçant. En pratique, on réalise donc pour les remplacements un chaînage partiel infra-annuel.

Dans le cadre des simulations réalisées, on adopte un principe simple de remplacement **fondé sur la marque**, qui représente les 6 à 11 premières positions du code-barres[22] (sur les 13 positions du code-barres, la dernière constituant une clé de contrôle calculée à partir des 12 premiers chiffres).

Lorsqu'un code-barres disparaît (c'est-à-dire lorsque ses ventes mensuelles sont nulles), on le remplace par un code-barres stable dans le même point de vente dont le préfixe est plus proche possible du code-barres disparu.

On teste ainsi l'existence dans le même point de vente d'un code-barres stable ayant les mêmes 11 premières positions. S'il en existe plusieurs, on prend celui dont les ventes du mois courant sont les proches de celles du code-barres disparu le mois précédent[23]. Sinon teste l'existence de code-barres ayant les 10 mêmes premières positions que le code-barres disparu, puis 9, 8 ,7 et 6.

On notera ainsi qu'un code-barres est remplacé au plus une fois, compte-tenu du fait que son remplaçant est choisi parmi les codes-barres stables, ce qui permet de simplifier les traitements[24].

Cette méthode permet de remplacer entre 65% et 85% des séries instables pour les principales familles de produits présentes dans les données de test.

S'il n'existe aucun code-barres stable dans le point de vente ayant les 6 mêmes premières positions que le code-barres disparu, alors on impute à la série l'évolution moyenne des prix de la famille dans le point de vente, calculée comme la moyenne géométrique des évolutions de prix mensuelles des codes-barres stables ou pour lesquels un remplaçant a été trouvé.

On dispose ainsi d'une table donnant un prix mensuel pour chaque série, ce qui permet ensuite de calculer un indice des prix à partir des évolutions de prix des séries sélectionnées dans chacun des 500 échantillons simulés.

[22] Plus la marque gère un nombre important de produits et plus son préfixe sera court afin de disposer d'un réservoir de codes-barres contenant son préfixe plus important.

[23] L'objectif de cette règle est de choisir pour remplaçant au sein des codes-barres de la même marque un code-barres dont l'importance des ventes est proche de celle du code-barres disparu.

[24] Malgré cette simplification, les programmes de remplacements des codes-barres sont extrêmement long à tourner, généralement sur plusieurs jours.

Une limite cependant est qu'un même code-barres peut intervenir plusieurs fois dans le calcul de l'indice, par exemple s'il a été sélectionné directement dans l'échantillon initial et s'il intervient en plus comme remplaçant d'un code-barres instable sélectionné dans l'échantillon initial.

Compte-tenu du fait que 30% des codes-barres sont instables, cette remarque conduira donc plutôt par prudence à retenir une taille d'échantillon minimale de 10 000 codes-barres pour les simulations, afin de s'assurer qu'on a au moins 5 000 codes-barres distincts.

3.1.b Le calcul des micro-indices

Rappel : Dans le cadre de la collecte enquêteurs, on calcule un micro-indice par variété et agglomération (indice varagglo) adapté au mode d'échantillonnage IPC (cible de prix à observer par variété et par agglomération). Dans le cas des variétés homogènes, le micro-indice est calculé au moyen de rapport de sommes de prix. Dans le cas de variétés hétérogènes, on calcule la moyenne géométrique des évolutions de prix.

Compte-tenu du mode de tirage des séries dans une base de sondage globale pour chaque famille IRI, on calcule ici directement **un indice global par famille IRI**.

Comme il s'agit de familles relativement larges et donc constituées de produits hétérogènes, on calcule ici un indice par **moyenne géométrique des évolutions de prix de l'ensemble des produits de l'échantillon tiré dans la famille**. Une autre raison pour le choix de la moyenne géométrique est que cet indice est bien adapté au chaînage infra-annuel effectué pour les remplacements.

3.2 Résultats des simulations effectuées

3.2.a Résultats obtenus au niveau poste

Afin d'obtenir des résultats significatifs, les résultats sont observés au niveau poste IPC.

1) Précision des indices obtenus

Les simulations montrent une précision satisfaisante des indices de postes obtenus (indices significatifs à 1% près sur le champ des enseignes participant au test[25]) pour des taux de sondage compris entre 1% des séries pour les postes les plus importants en termes de consommation des ménages et 2% des séries pour les postes les moins importants.

A titre d'exemple, on donne ici les résultats observés pour différentes tailles d'échantillons pour le poste étudié le moins important en terme de consommation (riz) et celui qui est le plus important (yaourts)

Riz (famille IRI 4206) :

Taille globale échantillon	Allocation tirée	Evolution annuelle 2009 moyenne	Ecart-type	Minimum	Q1	Q5	Q95	Q99	Maximum
10 000	350	-2,1%	0,58%	-3,9%	-3,4%	-3,0%	-1,1%	-0,6%	-0,3%
20 000	700	-2,1%	0,40%	-3,3%	-2,9%	-2,7%	-1,4%	-1,2%	-1,0%
50 000	1750	-2,1%	0,23%	-2,8%	-2,6%	-2,4%	-1,7%	-1,5%	-1,3%

[25] C'est-à-dire que si l'indice calculé pour un poste donné sur le champ des enseignes participant au test est de +3%, on peut dire qu'il est compris entre +2% et +4%.

On constate ainsi que pour un échantillon de global de 10 000 codes-barres (soit un taux de sondage de 1% des codes-barres), on obtient dans 90% des cas un indice des prix du riz, sur les 6 enseignes considérées, significatif à 1% près.

Lorsqu'on double la taille de l'échantillon, l'indice est significatif à 1% dans 98% des cas.

Yaourt (famille IRI 5701) :

Taille globale échantillon	Allocation tirée	Evolution annuelle 2009 moyenne	Ecart-type	Minimum	Q1	Q5	Q95	Q99	Maximum
10 000	1795	-4,4%	0,23%	-5,6%	-5,0%	-4,8%	-4,0%	-3,9%	-3,7%
20 000	3590	-4,4%	0,16%	-4,8%	-4,8%	-4,7%	-4,1%	-4,0%	-3,9%
50 000	8980	-4,4%	0,10%	-4,7%	-4,7%	-4,6%	-4,2%	-4,2%	-4,1%

Dans le cas de la famille Yaourt, un taux de sondage de 1% suffit à obtenir un indice significatif à 1% dans 98% des cas.

2) Comparaison avec les indices issus de la collecte IPC

Les indices calculés au niveau de chaque poste à partir des données de caisses sont comparés avec les indices du poste suivants issus de l'IPC :

- indice poste IPC calculé dans le cadre de la collecte actuelle (toutes enseignes et toutes formes de ventes, avec la technique des indices varagglo)
- indice « grande distribution » calculé toutes enseignes confondues sur les formes de ventes étudiées (hypermarchés, supermarchés et magasins populaires), comme la moyenne géométrique des évolutions de prix pour l'ensemble des produits IPC du champ appartenant au poste
- indice « grande distribution/enseignes du test » calculé sur les enseignes participant au test et les formes de ventes étudiées, comme la moyenne géométrique des évolutions de prix pour l'ensemble des produits IPC du champ appartenant au poste

On obtient les résultats suivants :

Riz

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	-2,1%
Indice IPC actuel	0%
Indice IPC grande distribution toutes enseignes confondues	-2,4% (286 observations)
Indice IPC grande distribution sur les enseignes du test	+1,3% (82 observations)

Huiles alimentaires

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	-5,9%
Indice IPC actuel	-5,3%
Indice IPC grande distribution toutes enseignes confondues	-4,7% (312 observations)
Indice IPC grande distribution sur les enseignes du test	-5,1% (91 observations)

Jus de fruits

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	+1,7%
Indice IPC actuel	+2,6%
Indice IPC grande distribution toutes enseignes confondues	+2,1% (241 observations)
Indice IPC grande distribution sur les enseignes du test	+0,2% (60 observations)

Chocolat en tablette

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	-0,1%
Indice IPC actuel	+0,2%
Indice IPC grande distribution toutes enseignes confondues	-0,8% (343 observations)
Indice IPC grande distribution sur les enseignes du test	+1,7% (103 observations)

Yaourts

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	-4,4%
Indice IPC actuel	-4,0%
Indice IPC grande distribution toutes enseignes confondues	-4,3% (666 observations)
Indice IPC grande distribution sur les enseignes du test	-5,7% (166 observations)

Cafés

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	+2,1%
Indice IPC actuel	+2,4%
Indice IPC grande distribution toutes enseignes confondues	+2,5% (535 observations)
Indice IPC grande distribution sur les enseignes du test	+1,1% (156 observations)

Œufs

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	-1,0%
Indice IPC actuel	-0,7%
Indice IPC grande distribution toutes enseignes confondues	-1,7% (449 observations)
Indice IPC grande distribution sur les enseignes du test	-2,6% (135 observations)

Fromages à pâte molle

	Evolution annuelle 2009
Indice simulé à partir des données de caisses	-2,4%
Indice IPC actuel	-3,0%
Indice IPC grande distribution toutes enseignes confondues	-2,8% (919 observations)
Indice IPC grande distribution sur les enseignes du test	-3,6% (295 observations)

On constate ainsi une bonne concordance entre les indices Données de Caisses et les indices grandes distribution toutes enseignes confondues (écart limité à 1%). Les divergences avec l'indices IPC calculé sur le seul champ des enseignes participant au test viennent avant tout des erreurs d'échantillonnage de l'IPC, compte-tenu du champ très restreint sur lequel portent les indices et du faible nombre de relevés de prix associés.

3.2.b Calculs d'indices de prix globaux sur les 8 postes étudiés

1. *Indice global portant sur les seules enseignes participant au test*

On étudie ici l'indice d'évolution globale des prix sur les 8 postes étudiés et sur les enseignes participant au test. On compare ici l'indice global issu des données de caisses (moyenne pondérée des indices de postes « données de caisses » calculés au paragraphe précédent) et l'indice issu de la collecte IPC (moyenne pondérée des indices de poste « IPC grandes distribution sur les enseignes du test »). Les pondérations sont les ventes totales annuelles 2008 des enseignes participant au test (estimées à partir des données de test). On obtient les résultats suivants :

Poste	**Poids du poste**	**Indices DDC**	**Indice IPC**	**Intervalle de confiance à 95% de l'indice IPC**	
Café	15,6%	2,1%	1,1%	*0,5%*	*3,7%*
Chocolat en tablette	11,8%	-0,1%	1,7%	*-1,8%*	*1,6%*
Huile	8,5%	-5,9%	-5,1%	*-8,2%*	*-3,6%*
Riz	3,8%	-2,1%	1,3%	*-5,8%*	*1,6%*
Yaourt	21,1%	-4,4%	-5,7%	*-5,9%*	*-2,9%*
Fromages pâte molle	15,6%	-2,4%	-3,6%	*-3,7%*	*-1,1%*
Œufs	9,9%	-1,0%	-2,6%	*-2,8%*	*0,8%*
Jus de fruits	13,6%	1,7%	0,2%	*0,2%*	*3,2%*
Ensemble	**100,0%**	**-1,4%**	**-2,0%**	***-2,0%***	***-1,1%***

La proximité de l'indice Données de Caisses (-1,4%) et l'indice IPC (-2,0%) est satisfaisante, compte-tenu du champ restreint sur lequel ils portent (environ 3% des ventes de la grande distribution) et donc de l'importance de l'erreur d'échantillonnage dans la collecte IPC.

Par ailleurs, il est possible avec les données de caisse de simuler un intervalle de confiance à 95% pour une taille d'échantillon égale à celle de la collecte IPC[26]. On constate que, sauf exception, l'indice IPC calculé à partir de la collecte enquêteurs tombe dans l'intervalle de confiance simulé à partir des données de caisse, ce qui valide le constat d'une proximité satisfaisante des deux indices compte tenu de la variance d'échantillonnage dans l'IPC actuel sur le champ de l'étude.

2. Indice global portant sur l'ensemble de la grande distribution

On calcule ici des indices portant sur l'ensemble du champ de la grande distribution. On compare un indice « purement IPC » calculé à partir des seuls relevés IPC avec un indice « mixte » calculé à partir des données de caisses pour les enseignes participant au test et à partir des relevés IPC pour les enseignes non participantes.

Les données de ventes annuelles 2008 « toutes enseignes confondues » sont issues des fichiers de consommation annuelle fournis à l'INSEE par une société d'études de marchés.

L'indice IPC « toutes enseignes confondues » est calculé comme la moyenne pondérée par les ventes totales annuelles 2008 « toutes enseignes confondues » des indices de poste « toutes enseignes confondues » calculés au paragraphe précédent.

Pour calculer l'indice mixte « Données de Caisse/IPC » global, on calcule d'abord un indice mixte au niveau de chaque poste, comme une moyenne pondérée de l'indice données de caisses pour les enseignes participant au test et de l'indice IPC calculé sur les enseignes ne participant pas au test.

Le poids de l'indice « données de caisses » dans l'indice mixte correspond à la part des enseignes participant au test dans les ventes totales de la grande distribution pour le poste considéré, calculée comme le rapport entre les ventes totales annuelles 2008 des enseignes participant au test (estimées à partir des données de test) et les ventes totales annuelles 2008 de la grande distribution (issues des données de consommation annuelle fournies par une société d'études de marchés)

L'indice mixte global « Données de Caisses/IPC » est ensuite calculé comme la moyenne des indices mixtes au niveau poste pondérée par les ventes annuelles 2008 « toutes enseignes confondues ».

On obtient les résultats suivants :

Poste	Indice IPC TOUTES ENSEIGNES	Indice Mixte « Donnes de caisse / collecte IPC »
CAFES TORREFIES	2,5%	2,8%
CHOCOLAT EN TABLETTE	-0,8%	-1,4%
HUILES	-4,7%	-4,9%
RIZ	-2,4%	-2,0%
YAOURT	-4,0%	-3,9%
FROM. PATE MOLLE	-2,8%	-2,5%
OEUFS	-1,7%	-1,3%
JUS DE FRUIT	2,1%	2,4%
Ensemble	**-1,5%**	**-1,3%**

On constate une bonne proximité entre l'indice IPC « toutes enseignes confondues » (-1,5%) et l'indice mixte (-1,3%), compte-tenu du fait que les 8 postes suivis couvrent de l'ordre de 10% des ventes de la grande distribution.

[26] Simulation sur la base de 500 tirages.

Conclusion

L'étude de faisabilité réalisée entre septembre 2009 et mars 2011 a permis de conclure sans ambiguïté de façon positive quant à la faisabilité méthodologique du projet données de caisse.

Sur le plan méthodologique, les travaux de simulations d'indices donnent des résultats satisfaisants quant à la précision des indices « données de caisse » et à leur comparaison avec les indices issus de l'IPC actuel.

L'étude de l'échantillon de données de test montre certes une forte instabilité des codes-barres[27] qui ne permettrait pas de prendre en compte l'ensemble des séries (croisement d'un code-barres et d'un point de vente) dans le calcul de l'indice, compte tenu de l'importance des remplacements à effectuer.

Cependant, les simulations de calcul d'indices de prix montrent qu'il est possible d'obtenir un niveau de précision satisfaisant des indices (indices de postes significatifs à 1% près sur le champ des six enseignes du test) avec des tailles d'échantillon réduites de l'ordre de 1% à 2% des séries**.**

D'autre part, la comparaison entre les indices « données de caisse » et les indices issus de la collecte enquêteurs montre une proximité satisfaisante entre ces indices, compte tenu notamment des erreurs d'échantillonnages de l'IPC sur le champ de l'étude.

[27] Taux de disparition des séries en cours d'année de 45%.