Projet d'utilisation des données de caisse dans le calcul de l'indice des prix





Sébastien FAIVRE
JMS 2012
25 janvier 2012

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

L'indice des prix à la consommation

- Un indicateur fondamental …
  - Mesure de l'inflation (érosion monétaire)
  - Déflateur des comptes nationaux (calcul des évolutions en volume : pouvoir d'achat, PIB …)
  - Indexation des contrats (IPC hors tabac)
  - Indicateur principal de la Banque centrale européenne
- … calculé par l'Insee
  - Un cadre international harmonisé et des règlements européens
  - Suivi d'un panier fixe de biens (évolution « pure » des prix) couvrant 95 % du champ de la consommation des ménages
  - Traitements « qualité » lors des renouvellements de produits
  - 180 000 relevés par mois dans 27 000 points de vente

Des demandes nouvelles pour les statistiques de prix

- IPC régionaux et comparaisons spatiales
- Suivi de segments fins : produits bio, éco-labellisés … (développement durable)
- Mieux tenir compte des nouveaux modes de vie pour approcher le coût de la vie (calcul de prix moyens, i.e. ceux des téléviseurs, des ordinateurs …) (rapport Quinet)
- Comparaisons des niveaux de prix entre pays européens (tableau de bord du grand marché intérieur de la DG Santé et protection du consommateur)

La segmentation et complexification des marchés de consommation

- Du côté de l'offre, les marchés de consommation se diversifient et/ou se complexifient de plus en plus :
  - 470 000 références pour les produits de grande consommation hors vins : 1 600 nouvelles références chaque semaine
  - Multiplication des segments : produits à bas coût, produits diététiques, bio, hallal …
  - Multiplication des promotions (10% du CA pour les produits de grande consommation)
  - Prix quasi personnalisés (billets train ou d'avion)
  - Tarifications au forfait (téléphonie, services bancaires, …)

⇒Ces tendances questionnent la représentativité des paniers de consommation de l'IPC ou compliquent l'observation des prix

⇒L'Insee accède de plus en plus aux bases ou données professionnelles : médicaments (base ISMHEALTH : tous médicaments de 60% des pharmacies), médecins/dentistes (Cnam), billets d'avion (base DGAC), billets de train (SNCF), services bancaires (FFB), téléphonie mobile (enquête auprès des opérateurs privés puis Arcep), assurances

Les limites de l'indice des prix à la consommation actuel

- Un échantillon d'agglomérations datant de 1990
- Des méthodes datées de traitement des promotions et des remplacements
- Un nombre limité de variétés (un millier) et de séries élémentaires dans un contexte de segmentation des marchés et de fort renouvellement des produits

⇒ **L'IPC doit évoluer**

Des innovations à l'étranger

- Des initiatives étrangères à partir des relevés internet :
  - Google Price Index
  - Billion Price Index (MIT)
- Des expériences réussies d'exploitation des données de caisse dans plusieurs pays
- Projets de moyen terme d'Eurostat : beaucoup de pays européens réfléchissent à l'exploitation des données de caisse
- Les données de caisse sont le sujet majeur de la réflexion internationale et de la recherche académique depuis 10 ans

L'exploitation des données de caisse : un gain d'information considérable

- Une décision stratégique : reproduire la méthode actuelle de l'indice des prix, seule méthode éprouvée actuellement et qui assure la continuité des séries :
  - Indice de Laspeyres, chaîné annuellement
- Des données de caisse : un gain d'information considérable
  - Données brutes des ventes du jour (prix, quantités) : prix affichés en magasin et qui figurent sur les tickets remis à la caisse aux clients
  - Base de sondage exhaustive (connaissance de l'univers) : calcul de précision …
  - Connaissance des prix et quantités (constitution de paniers annuels représentatifs, repérages des promotions, traitement des remplacements)

⇒Des études et simulations seront conduites sur la bonne utilisation des données pour le partage prix/qualité

Un saut qualitatif majeur de la statistique de prix

Les données de caisse permettront de :

–Répondre à des demandes nouvelles : comparaisons spatiales de prix, suivi de marchés particuliers (bio …), prix moyens, demandes européennes

–Améliorer la précision de l'IPC et la représentativité de son échantillon (augmentation du nombre de séries de prix suivies)

–Homogénéiser les traitements qualité et ceux relatifs aux promotions

LES ASPECTS METHODOLOGIQUES DE L'UTILISATION DES DONNES DE CAISSE DANS LE CALCUL DE L'INDICE DES PRIX

Plan de la présentation

I. Présentation du système cible

II. Les travaux méthodologiques menés sur les données de caisse

III. La gestion des remplacements et l'estimation de l'effet qualité

IV. Les conditions de réalisation du projet

- faisabilité de la collecte restante
- mise en place d'un plan de secours

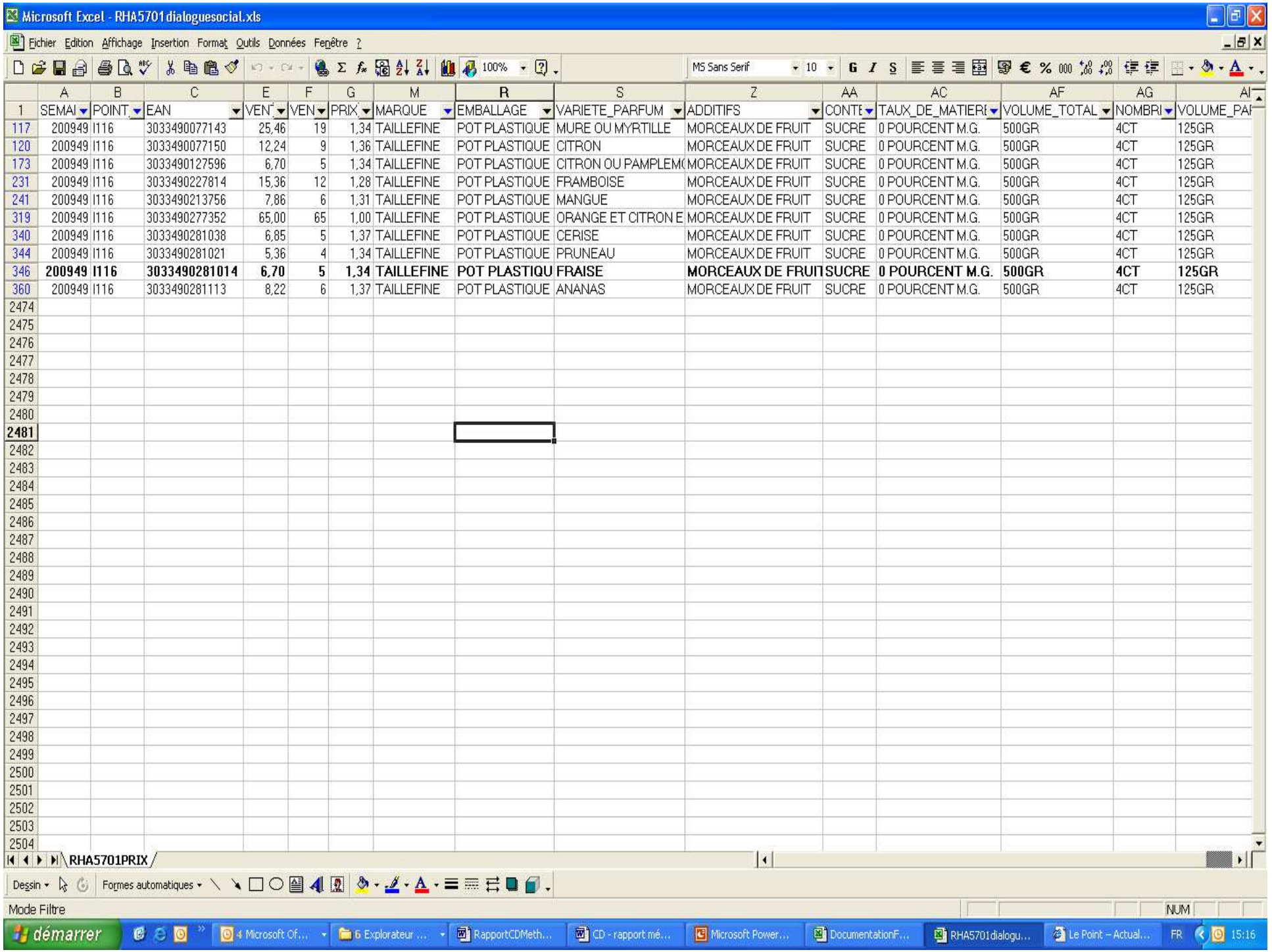
I. Présentation du système cible

Les données à collecter

Données observées aux caisse dans la quasi-totalité de supermarchés et d'hypermarchés : plusieurs dizaines de millions de prix relevés chaque jour

Enrichies par un répertoire de codes-barres

Champ restreint aux grandes et moyennes surfaces (supermarchés et hypermarchés). Exclusion des supérettes (surface de vente inférieure à 400 m2) et du hard discount qui ne remontent pas systématiquement de données de caisse



	A	B	C	E	F	G	M	R	S	Z	AA	AC	AF	AG	AI
1	SEMAI	POINT	EAN	VEN	VEN	PRIX	MARQUE	EMBALLAGE	VARIETE_PARFUM	ADDITIFS	CONTE	TAUX_DE_MATIERE	VOLUME_TOTAL	NOMBRI	VOLUME_PAI
117	200949	I116	3033490077143	25,46	19	1,34	TAILLEFINE	POT PLASTIQUE	MURE OU MYRTILLE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
120	200949	I116	3033490077150	12,24	9	1,36	TAILLEFINE	POT PLASTIQUE	CITRON	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
173	200949	I116	3033490127596	6,70	5	1,34	TAILLEFINE	POT PLASTIQUE	CITRON OU PAMPLEM(	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
231	200949	I116	3033490227814	15,36	12	1,28	TAILLEFINE	POT PLASTIQUE	FRAMBOISE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
241	200949	I116	3033490213756	7,86	6	1,31	TAILLEFINE	POT PLASTIQUE	MANGUE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
319	200949	I116	3033490277352	65,00	65	1,00	TAILLEFINE	POT PLASTIQUE	ORANGE ET CITRON E	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
340	200949	I116	3033490281038	6,85	5	1,37	TAILLEFINE	POT PLASTIQUE	CERISE	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
344	200949	I116	3033490281021	5,36	4	1,34	TAILLEFINE	POT PLASTIQUE	PRUNEAU	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR
346	**200949**	**I116**	**3033490281014**	**6,70**	**5**	**1,34**	**TAILLEFINE**	**POT PLASTIQU**	**FRAISE**	**MORCEAUX DE FRUIT**	**SUCRE**	**0 POURCENT M.G.**	**500GR**	**4CT**	**125GR**
360	200949	I116	3033490281113	8,22	6	1,37	TAILLEFINE	POT PLASTIQUE	ANANAS	MORCEAUX DE FRUIT	SUCRE	0 POURCENT M.G.	500GR	4CT	125GR

Un nombre important de produits suivis

Nombre moyen de références par magasin (codes-barres) en mai 2010 dans les hypermarchés et les supermarchés (source IRI)

Quelques exemples :

Produit	Nombre moyen de références par magasin
Tablettes de chocolat	159
Pâtes alimentaires sèches	150
Jus de fruits	125
Chewing Gum	74

Une documentation complète des codes-barres

Exemple : pâtes alimentaires

- Segment (pâtes supérieures, pâtes aux œufs, pâtes farcies, pâtes farcies aux œufs, pâtes exotiques)
- Fabricant
- Marque
- Nom
- Mode de conditionnement
- Poids du paquet

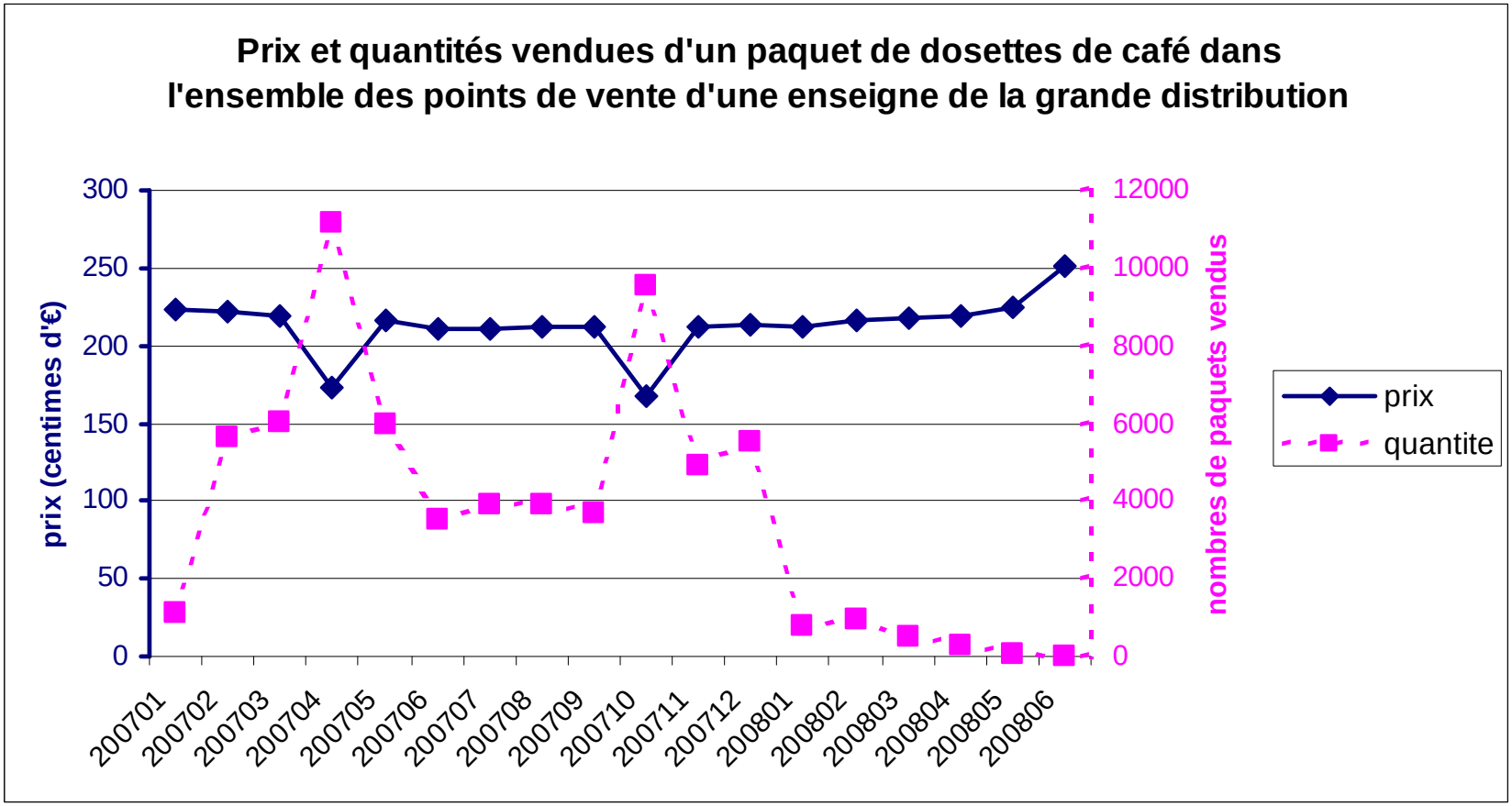
…

Une connaissance beaucoup plus fine de la consommation

Connaissance également des quantités vendues

- Possibilité de sélectionner des paniers annuels « représentatifs » de la consommation des ménages proportionnellement aux ventes

- Possibilité d'observer « en continu » le prix moyen des articles vendus pour une famille de produits donnée (impact important des promotions) et de mener des travaux de recherche sur des indices de prix moyens


Prix et quantités vendues d'un paquet de dosettes de café dans l'ensemble des points de vente d'une enseigne de la grande distribution
prix (centimes d'€)
nombres de paquets vendus
300
250
200
150
100
50
0
12000
10000
8000
6000
4000
2000
0
200701
200702
200703
200704
200705
200706
200707
200708
200709
200710
200711
200712
200801
200802
200803
200804
200805
200806
prix
quantite

Les questions méthodologiques à étudier

a. Le code-barres : un identifiant statistique du produit

Le code-barres est attribué par le fabricant pour le suivi du produit à toutes étapes de la chaîne de distribution

Il permet un repérage et un suivi des produits beaucoup plus précis et plus fiable que dans l'IPC actuel pour les produits de fabrication industrielle

b. Un renouvellement rapide des codes-barres

Etude norvégienne sur le champ de l'alimentaire (hors boissons alcoolisées) pour l'année 2004:

- au bout d'un mois, 27% des codes-barres ont disparu
- au bout d'un an, 58% des codes-barres ont disparu

=> Importance des remplacements à effectuer du fait de la disparition des produits suivis

On dispose en revanche de beaucoup plus d'information qu'actuellement (y c. fonction de demande) pour les remplacements et pour estimer les effets qualité (qui seront étudiés par le projet)

c. La question des ventes irrégulières pour les biens durables

Il s'agit ici de données de caisse : si le produit n'est pas vendu un mois donné, on n'a aucune information sur l'évolution du prix

Cela peut poser problème pour le suivi du prix de certains biens durables, comme les cafetières par exemple

II. Les travaux méthodologiques menés sur les données de caisse

(premiers résultats provisoires)

Les données de test commandées

Commande de données de test auprès de la société IRI: données de ventes hebdomadaires portant sur 10 familles de produits, 3 années (2007 à 2009) et 1000 points de ventes

10 familles de produits de grande consommation représentant au total 1000 codes-barres en moyenne par magasin

Elles correspondent à huit postes de l'IPC actuel : café torréfié, chocolat en tablette, jus de fruits, œufs, yaourt, huiles alimentaires, riz, fromages à pâte molle

Plusieurs enseignes ont donné leur accord pour l'acquisition de données de test

Ces enseignes représentent environ 30% de la surface de vente totale de la grande distribution alimentaire

Données reçues début novembre 2010

Les objectifs du test méthodologique

- Comparer les données de caisse avec celles issues de la collecte enquêteurs

- Etudier la stabilité dans le temps des séries élémentaires (croisement code-barres*point de ventes)

- Simuler des calculs d'indices de prix à partir des données scannées à méthode inchangée

- Comparer les indices obtenus avec les indices issus de la collecte enquêteurs

a. Le test de comparaison entre les données de caisse et la collecte IPC

Première étape : pour chaque produit suivi par un enquêteur, identifier dans les données de caisse le ou les codes-barres pouvant correspondre à ce produit, sur la base des seules caractéristiques techniques du produit

Travail effectué par trois sites prix volontaires (Lille, Lyon et Nancy) sur 345 produits relevés en décembre 2009 et rentrant dans le champ de l'étude

Puis vérification de la correspondance entre le prix relevé par l'enquêteur et les prix moyens hebdomadaires des codes-barres appariés au produit

Résultats : taux d'appariement de 75%

60% des produits appariés avec un code-barres ayant un prix identique au prix relevé par l'enquêteur et **15%** avec un prix approchant (différence tolérée de 5% compte tenu du fait qu'on compare le prix relevé un jour donné par l'enquêteur avec un prix moyen dans les données de caisse

Une étude complémentaire a été menée par le site prix de Lyon sur les produits non appariés.

3 causes principales:

- produits absents en rayon lors du passage de l'enquêteur (prix de novembre 2009 reconduit)
- difficultés de mise à jour par les enquêteurs des caractéristiques techniques des produits
- difficultés spécifiques de collecte à la veille de Noël

Ce test a mis en évidence la qualité des données de prix et de la documentation des codes-barres issues des données de caisse

Les données de caisse permettraient ainsi d'augmenter fortement la qualité de l'indice en disposant de relevés de prix exhaustifs et à fréquence journalière avec une documentation complète et à jour des codes-barres

b. Etude de la stabilité dans le temps des séries

Unité élémentaire : série (croisement code-barres et point de vente)

Base de sondage pour le calcul de l'indice des prix 2009 à partir des données de test (10 familles de produits et 1000 points de vente appartenant aux enseignes participantes) : **1 096 604** séries présentes au mois de base (décembre 2008)

Les premiers résultats du test méthodologique

Principale difficulté pour le suivi d'un panier annuel : l'instabilité des codes-barres

On appelle ici code-barres **stable** dans un point de ventes un code-barres dont les ventes sont strictement positives tous les mois de l'année 2009 (dans le cas contraire, si au moins une des ventes mensuelles est nulle, le code-barres est instable)

Les premiers résultats du test méthodologique

Sur la base de sondage étudiée, un peu plus de la moitié des séries sont stables (55%)

Près de 500 000 remplacements à effectuer sur le champ étudié si on retenait l'ensemble des séries (3% de la consommation)=> 15 millions de remplacements annuels

Les remplacements seront autant que possible automatisés ...

…Mais une automatisation totale des remplacements ne sera peut-être pas possible!

=> Nécessité de se limiter au suivi d'un échantillon de codes-barres

La sélection d'un échantillon de séries

Les séries stables représentent 72% des ventes annuelles (CA)

=> sélection d'un échantillon de séries proportionnellement aux ventes, pour cibler les codes-barres bien vendus qui sont plus stables que la moyenne

On atteindrait alors un taux annuel de remplacement des séries de 28%, contre 17% actuellement dans l'IPC sur les 10 familles de produits étudiées

Cette différence s'explique en partie par une meilleure prise en compte des promotions « fabricant » dans l'indice (dans les données de test, les promotions « fabricant » représentent 7% des codes-barres et de l'ordre de 10% des ventes)

c. Simulation d'indices de prix à partir des données de caisse

Simulations de l'inflation annuelle 2009 (évolution des prix entre décembre 2008 et décembre 2009) sur les huit postes de l'indice des prix présents dans les données de test :

- taux de sondage de 1%, 2% et 5% (0,2% dans l'IPC actuel) : échantillons de 10, 20 et 50 000 codes-barres
- 500 tirages d'échantillon avec équilibrage sur la marque et sur l'enseigne
- mise en œuvre de la technique du « recouvrement » pour les remplacements (on connaît le prix du produit remplaçant au mois de base

Méthodologie de calcul de l'indice

Remplacement des codes-barres effectués de manière automatique en recherchant un code-barre stable de la même marque (même préfixe que le code-barres à remplacer) dans le même point de vente.

Si aucun code-barres remplaçant n'est trouvé, on impute l'évolution moyenne des prix dans le point de vente

On calcule ensuite l'indice d'évolution comme la moyenne géométrique des évolutions de prix des séries de l'échantillon

d. Le résultat des simulations

Précision des indices obtenus à partir des données de caisse

Objectif: obtenir au niveau poste un indice significatif à un point de pourcentage

Les simulations montrent qu'on peut atteindre une précision satisfaisante pour des tailles d'échantillons réduites : taux de sondage de 1% à 2% des séries.

Comparaison avec les indices issus de l'IPC

- Indice IPC du poste « toutes formes de ventes confondues » préexistant

Deux indices complémentaires :

- Indice IPC du poste « grande distribution, toutes enseignes confondues » et indice IPC « grandes distribution, enseignes du test », calculés comme la moyenne géométrique des évolutions des prix relevés par les enquêteurs sur les produits du champ

Résultat de la comparaison

Sur les huit postes, les indices « données de caisse » calculés sur les enseignes du test sont quasiment tous proches des indices IPC « grandes distribution, toutes enseignes confondues » (écart inférieur à 1%)

Les différences sont plus grandes avec les indices IPC portant seulement sur les enseignes du test, compte tenu du faible nombre d'observations dans l'IPC et de l'aléa d'échantillonnage associé.

Exemples: Indices d'inflation annuelle 2009 pour les familles Yaourt et Riz

	Yaourt	Riz
Indice calculé à partir des données de caisse	-4,4%	-2,1%
Indice IPC toutes formes de ventes confondues	-4,0%	0%
Indice IPC « grande distribution, toutes enseignes »	-4,3%	-2,4%
Indice IPC « grande distribution, enseignes du test »	-5,7%	+1,3%

Calcul d'un indice global sur les huit postes suivis et les six enseignes du test

Indice global d'inflation annuelle 2009

Indice estimé à partir des données de caisse: -1,4%

Indice estimé à partir de la collecte enquêteurs: -2,0%

Intervalle de confiance à 95% pour l'indice issu de la collecte enquêteurs: [-2,0%,-1,1%]

=> **les différences observées sont liées à l'aléa d'échantillonnage de l'IPC**

Conclusion

L’utilisation des données de caisse permettrait d’améliorer fortement la précision des indices à un niveau fin, même avec des tailles d’échantillon réduites (1 à 2% des séries) compatibles avec la gestion des remplacements

Conclusion

Il s’agit à ce stade de premiers résultats provisoires

Ces résultats seront affinés et complétés dans le cadre des travaux méthodologiques menés par le CPS adjoint données de caisse au cours de l’année 2012