Echantillon optimal d'agglomérations pour l'IPC français après incorporation des données de caisse





Sébastien FAIVRE
INSEE
Colloque Sondages 2012, ENSAI

INSEE
INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

08/06/2012

Plan de la présentation

I. Présentation de l'IPC et du projet données de caisse

II. Définition de l'échantillon théorique optimal d'agglomérations pour la collecte des prix restante après incorporation des données de caisse

I. Présentation de l'IPC et du projet données de caisse

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

Le plan de sondage de l'IPC en métropole

Tirage à deux degrés

Premier degré: tirage des agglomérations de relevés

Deuxième degré: tirage des produits suivis dans les agglomérations

Le plan de sondage des prix observés pour l'IPC: premier degré

Tirage de 96 agglomérations, stratifié par ZEAT (groupes de régions) et par tranche de taille d'unités urbaines

	Tranche de taille d'unités urbaines (au RP 1990)	Nombre d'agglomérations de l'échantillon
A	Paris	1
B	Plus de 100 000 habitants	37
C	Entre 20 000 et 100 000 habitants	25
D	Moins de 20 000 habitants	33

Le plan de sondage des prix observés pour l'IPC: deuxième degré de tirage

La consommation des ménages est suivi au travers de 1000 variétés, par exemple:

- **Gaz butane comprimé (13 kg), sans consigne**
- **Pain baguette (Prix au kg)**
- **Pain parisien (Prix au kg)**
- **Camembert pasteurisé 45% Matière Grasse fabriqué en Normandie (Prix des 250g)**
- **Lait UHT demi - écrémé (Prix du litre)**
- **Escalope de veau (Prix au kg)**
- **Café moulu 100% arabica (Prix des 250g)**

Pour chaque variété, un tableau d'équilibrage donne le nombre de produits à suivre par variété et par forme de ventes

Exemple de tableau d'équilibrage pour une variété

AGGLO	HYPER-MARCHE	SUPER-MARCHE	MAXIDIS COMTE	SUPERETTE	MAG.N POPU-LAIRE	AUTRE			MARCHE	TOTAL
1	2	2	1	0	0	…	…	…	1	6
2	1	2	1	0	0	…	…	…	0	4
3	1	1	1	1	1	…	…	…	0	5
…..	….	….	….	….	….	…	…	…	….	….
…..	….	….	….	….	….	…	…	…	….	….
15	8	6	3	1	1	0	…	1	1	21
…..	….	….	….	….	….	…	…	…	….	….
…..	….	….	….	….	….	…	…	…	….	….
24	2	1	1	0	0				0	4
TOTAL	60	52	20	5	5	0	…	4	4	150

Le plan de sondage des prix observés pour l'IPC: deuxième degré de tirage

Dans ce cadre, les enquêteurs choisissent les produits suivis dans les magasins, en cherchant également à diversifier les produits suivis (type de marque, produits biologiques ou conventionnels…) de façon à refléter la diversité de la consommation des ménages

La production de l'IPC par l'INSEE

- Une mission majeure de l'Insee
  150 enquêteurs, 9 sites prix (70 ETP et 6 A), division des prix (20 personnes)

- Un cadre international harmonisé et des règlements européens

- Les limites de l'IPC actuel :
  - Le traitement des promotions et des remplacements
  - Un nombre limité de variétés (un millier) et de séries élémentaires
  - Un échantillon d'agglomérations datant de 1990

⇒ Nécessité d'une évolution de l'IPC

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

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

- Les 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 (fonction de demande par magasin : élasticités-prix et de substitution entre produits, repérages des promotions, traitement des remplacements)

- Des expériences réussies d’exploitation des données de caisse dans plusieurs pays

⇒Les premiers travaux méthodologiques pour l’utilisation des données de caisse dans l’IPC ont été présentés aux JMS 2012

Un impact important sur le nombre de relevés à effectuer

Les relevés de prix effectués sur le champ données de caisse (440 000 relevés par an) représentent un tiers des relevés de prix effectués chaque année par les enquêteurs (1 344 000).

Cependant, tous les relevés ne correspondant pas au même temps de relevé: par exemple, un relevé dans l'habillement prend environ moitié plus de temps en moyenne qu'un relevé dans l'alimentaire

En terme de temps de relevés, les données de caisse couvrent environ un quart de la charge de travail totale des enquêteurs

Une réorganisation de la collecte à mettre en œuvre, en conservant la précision de l'indice

Compte tenu de la baisse importante du volume de collecte, une réorganisation de la collecte restante doit être envisagée

L'hypothèse théorique étudiée ici est celle d'une concentration de la collecte restante dans un nombre restreint d'agglomération de façon à maintenir une charge de travail suffisante par enquêteur

L'objectif est de vérifier que, suite à cette concentration de la collecte, on conserve pour l'indice d'ensemble une précision au moins égal à la précision actuelle

II. Définition de l'échantillon d'agglomération optimal pour la collecte des prix restante après incorporation des données de caisse

Le champ de l'étude

On se contentera ici d'étudier la précision **sur le champ couvert par la collecte enquêteurs**, puisque le champ « hors collecte enquêteurs » (tarifs) n'est pas impacté par le passage aux données de caisses

De plus, l'hétérogénéité des sources utilisées pour les tarifs (données opérateurs pour la téléphonie mobile, données panels privés pour les médicaments, données DGAC pour le transport aérien…) rend difficile d'effectuer un calcul de précision sur le domaine « hors collecte enquêteurs ».

Méthode de calcul des indices élémentaires

L'indice élémentaire est l'indice var-agglo calculé à partir des prix relevés pour une variété et une agglomération données comme un rapport de prix moyens (variétés homogènes) ou la moyenne géométrique des évolutions de prix (variétés hétérogènes) par rapport au mois de base (décembre N-1)

L'indice des prix d'une variété donnée au sein d'une tranche d'agglomération et d'une ZEAT données est calculé comme la moyenne arithmétique des indices var-agglo de la variété pour les agglomérations appartenant à la strate considérée :

$$\hat{I}(cc,z,v) = \frac{1}{m(cc,z,v)} \sum_{i=1}^{m(cc,z,v)} \hat{I}(i,v)$$

où $\hat{I}(i,v)$ est l'indice var-agglo de la variété v dans l'agglomération i et $m(cc,z,v)$ le nombre d'agglomérations de la strate dans lesquelles la variété est observée.

Le modèle d'Ardilly et Guglielmetti pour le calcul de la précision de l'IPC

Travaux effectués pour l'optimisation de l'échantillon d'agglomération actuel (base 1990)

On modélise le tirage des agglomérations comme un sondage aléatoire simple global au sein de chaque tranche d'agglomérations (on ne fait donc pas intervenir la dimension géographique « ZEAT », pour obtenir des résultats plus robustes).

Le deuxième degré (tirage des points de ventes au sein des agglomérations) est modélisé comme un sondage aléatoire simple, pour lequel on néglige la correction de population finie.

Enfin, on fait l'hypothèse que les échantillons de prix observés pour des deux variétés distinctes sont « indépendants ». On néglige ainsi l'effet de grappe lié aux points de ventes.

Le modèle d'optimisation d'Ardilly et Guglielmetti

Pour le changement de base 1990, on cherche à minimiser le total nombre d'agglomérations de collecte sous la contrainte de conserver le même niveau de précision que dans l'échantillon antérieur.

Sous l'hypothèse que la variance de second degré dépend essentiellement du nombre de relevés effectués et ne dépend donc pas du tirage des agglomérations, cela revient à conserver la même variance de premier degré que dans l'échantillon antérieur.

Compte-tenu de l'hypothèse de sondage aléatoire simple pour le tirage des agglomérations, la variance de première phase pour l'indice de la variété dans la strate définie par la tranche d'agglomération cc et la ZEAT z est

$$V^{1P}(\hat{I}(cc,z,v)) = \frac{1}{m(cc,z,v)}(1-\frac{m(cc,z,v)}{M(cc,z)})S^2(cc,z,v)$$

où $S^2(cc,z,v)$ est la « vraie »dispersion des indices var-agglo pour la variété-strate , $M(cc,z)$ le nombre total d'agglomérations dans la strate cc,z et $m(cc,z,v)$ le nombre d'agglomérations de l'échantillon dans lesquelles la variété v est observée.

Un estimateur « naïf » de cette quantité est obtenu en remplaçant la vraie dispersion $S^2(cc,z,v)$ par son estimation à partir de l'échantillon $s^2(cc,z,v)$

L'indice global s'écrit $$\hat{I} = \sum_{cc,z,v} w(cc,z,v)\hat{I}(cc,z,v)$$

Par indépendance des tirages variétés/strates

$$V^{1P}(\hat{I}) = \sum_{cc,z,v} w^2(cc,z,v)V^{1P}(\hat{I}(cc,z,v))$$

$$V^{1P}(\hat{I}) = \sum_{cc,z,v} w^2(cc,z,v)(1-\frac{m(cc,z,v)}{M(cc,z)})\frac{S^2(cc,z,v)}{m(cc,z,v)}$$

$$\hat{V}^{1P}(\hat{I}) = \sum_{cc,z,v} w^2(cc,z,v)(1-\frac{m(cc,z,v)}{M(cc,z)})\frac{s^2(cc,z,v)}{m(cc,z,v)}$$

Le programme d'optimisation s'écrit alors

$$Min \sum_{cc,z} m(cc,z)$$

Sous contrainte

$$\sum_{cc,z,v} w^2(cc,z,v)(1-\frac{m(cc,z)}{M(cc,z)})\frac{s^2(cc,z,v)}{m(cc,z)} = \hat{V}^{1P}_{antérieur}$$

Précision de l'IPC sur le champ de la collecte enquêteurs dans la situation actuelle

Pour la collecte actuelle, la variance totale est la somme de la variance de première phase et de la variance de seconde phase calculées selon la modélisation précédente.

Cette variance totale peut être estimée à partir de l'échantillon de prix relevés.

Elle constitue la valeur cible pour la variance après passage aux données de caisse

Précision de l'IPC sur le champ de la collecte enquêteurs avec données de caisse

La variance est composée de deux parties:

- variance sur le champ données de caisses: variance de premier degré plus variance de second degré

- variance sur le champ hors données de caisse: variance de premier degré plus variance de second degré

Modélisation du gain en variance associé aux données de caisse

Sur le champ couvert par les données de caisse :

Au premier degré: les données de caisse couvrent l'ensemble du territoire, donc la variance de premier degré est nulle

Au deuxième degré, la variance dépend du nombre de produits qui seront inclus dans le panier (nombre limité du fait de la gestion des remplacements). On anticipe néanmoins une très forte hausse du nombre de séries suivies sur la champ données de caisse (multiplication par 20).

On travaille ici avec deux scénarios: un scénario optimiste où la variance de second degré sur le champ données de caisse tombe à 0 (hypothèse **H0**) et un scénario prudent où la variance de second degré est divisée par deux (hypothèse **H2**).

Le calcul de la variance de second degré sur le champ hors données de caisse

Comme Ardilly et Guglielmetti, on fait ici l'hypothèse que la variance de seconde phase dépend uniquement du nombre total de relevés effectués pour les variétés qui sont dans le champ de la collecte, mais pas du nombre d'agglomérations dans lesquelles la collecte est réalisée.

Comme on pense conserver le même nombre de relevés que dans la collecte actuelle sur le champ « hors données de caisses » dans la collecte future, on peut donc considérer que la variance de seconde phase sur le champ « hors données de caisses » est la même avant et après le passage aux données de caisses

Le calcul de la variance de second degré sur le champ hors données de caisse

Il s'agit là a priori d'une hypothèse prudente, puisque, en concentrant la collecte du champ « hors données de caisses » dans un nombre plus faible d'agglomérations à nombre de relevés constants, on augmente le nombre moyen de relevés par agglomérations et donc on diminue la variance de seconde phase.

Le calcul de la variance de second degré sur le champ hors données de caisse

On dispose alors d'une estimation de la variance de seconde phase sur le champ hors données après passage aux données de caisse en estimant à partir de l'échantillon de prix relevés la variance de seconde phase actuelle sur le champ hors données de caisse.

Cible de variance de premier degré sur le champ hors données de caisse

On en déduit la cible de variance de premier degré pour la collecte restante (sur le champ hors données de caisse), en retirant à la variance totale actuelle la variance sur le champ données de caisse et la variance de second degré sur le champ hors données de caisse

$$V^{1}_{Après,HorsDDC} = V^{Total}_{Actuel} - V^{Total}_{Après,DDC} - V^{2}_{Après,HorsDDC}$$

Pour des raisons de robustesse, on calcule uniquement un nombre optimal d'agglomération par tranche d'unités urbaines (sans tenir compte des ZEAT)

Le programme d'optimisation s'écrit alors

$$Min \sum_{cc} m(cc)$$

Sous contrainte

$$\sum_{cc,z,v} w^2(cc,z,v)(1-\frac{m(cc,z)}{M(cc,z)})\frac{s^2(cc,z,v)}{m(cc,z)} = \hat{V}^{1P}_{Après,HorsDDC}$$

En notant $g^2(cc) = \sum_z \sum_v w^2(cc,z,v) s^2(cc,z,v)$

On obtient $$m(cc) = \frac{g(cc)\sum_{CC} g(cc)}{\hat{V}^{1P}_{Après,HorsDDC} + \sum_{CC} \frac{g^2(cc)}{M(cc)}}$$

On compare ensuite l'échantillon d'agglomérations ainsi obtenu avec un échantillon « terrain » obtenu de la façon suivante:

- Pour les agglomérations de plus de 100 000 habitants (où plusieurs enquêteurs réalisent la collecte), on conserve toutes les agglomérations quitte à réduire le nombre d'enquêteurs

- pour les agglomérations de moins de 100 000 habitants (où un seul enquêteur réalise la collecte), on concentre la collecte dans un nombre restreint d'agglomérations de manière à maintenir la charge de travail médiane par enquêteur dans la tranche d'agglomération

Résultats de l'optimisation

Tranche d'agglomération (hors agglomération parisienne)	Echantillon actuel	Echantillon optimal H2	Echantillon optimal H0	Echantillon terrain
Plus de 100 000 habitants	37	36	31	37
Entre 20 000 et 100 000 habitants	25	12	10	17
Moins de 20 000 habitants	33	18	16	20

Conclusion

L'échantillon issu de la concentration de la collecte restante est supérieur dans les trois tranches d'unités urbaines non exhaustives à l'échantillon minimal nécessaire pour conserver la précision globale de l'indice sur le champ de la collecte enquêteurs

Il est donc possible de concentrer la collecte des prix restante dans un nombre restreint d'agglomérations sans dégrader la précision de l'indice d'ensemble

Il s'agit cependant là uniquement d'un travail théorique exploratoire, réalisé avant la mise en place des Nouvelles Conditions d'Emploi des Enquêteurs