Échantillonnage des agglomérations de l'IPC pour la base 2015

Patrick Sillard et Laurence Jaluzot

INSEE – France

Mars 2015

Introduction : les principes de l'IPC sur la métropole

- Fondé sur le suivi mensuel des prix de 120 000 produits observés sur le terrain et de 100 000 "tarifs"
- Appartenant à de 1 000 type de produits (les "variétés", notées v ci-après)
- Vendus dans des points de vente d'une centaine d'agglomérations (notées a ci-après)
- Tous les mois, les enquêteurs retournent dans les mêmes points de vente observer les mêmes produits
- Chaque produit contribue à l'IPC à travers l'évolution de son prix observée depuis le mois de décembre précédent
- De l'ensemble, on calcule un indice de prix de Laspeyres à panier fixé sur une année, chaîné annuellement

Le changement de base IPC

- La base actuelle date de 1998
- Nécessité de revoir la liste des regroupements élémentaires de diffusion suite à l'adoption de la nouvelle nomenclature internationale COICOP
- Nécessité de revoir la méthodologie de suivi des produits frais qui n'est plus conforme aux règlements européens
- Nécessité de rafraîchir l'échantillon d'agglomérations métropolitaines dans lesquelles la collecte est réalisée, l'actuel étant issu d'une sélection opérée sur la base du RP 1990

Plan

Plan

L'agrégation de Laspeyres

A est une partition de la géographie du territoire métropolitain, V l'ensemble de variétés de produits :

$$I = \sum_{a \in A, v \in V} w_{a,v} I_{a,v} = \sum_{v \in V} w_v \underbrace{\sum_{a \in A} w_a I_{a,v}}_{I_v}$$

où $I_{a,v}$ est l'indice de prix pour la variété-agglomération (varagglo) (a, v), w_v est le poids de la variété v dans la dépense de consommation des ménages et w_a est le poids de l'agglomération a dans la dépense de consommation des ménages. Autrement dit :

$$\sum_{a \in A} w_a = 1 \text{ et } \sum_{v \in V} w_v = 1$$

Estimation

$$I_v = \sum_{a \in A} w_a I_{a,v}$$

On peut estimer I par...

$$\hat{I}_v = \sum_{a \in \mathscr{A}} \omega_a I_{a,v}$$

où $\mathscr{A}$ est un échantillon tiré dans A et ω_a est un poids de sondage calculé en conformité avec :

1. le plan de sondage (probabilité d'inclusion de a dans $\mathscr{A}$) ;
2. le poids w_a que a devrait avoir dans la valeur exacte de I_v (poids en dépense de consommation des ménages).

Un sondage sur la dimension géographique

Le sondage sur la dimension géographique (sélection d'agglomérations)

- on connaît l'univers
- les coûts de collecte sont directement liés à cette dimension

Minimisation de la variance d'échantillonnage

$\hat{I}_v = \sum_{a \in \mathscr{A}} \omega_a I_{a,v}$ est un estimateur

de la somme de la variable $y_a = w_a I_{a,v}$. Cette variable est approximativement proportionnelle à w_a ($I_{a,v}$ peu différent de 1). Ainsi, la probabilité d'inclusion devrait idéalement être proportionnelle à w_a.

Stratification

La stratification est un moyen supplémentaire d'améliorer la précision.

Stratification : taille optimale des strates

- On définit des strates géographiques correspondant au croisement de 7 régions et de 4 types d'agglomérations regroupées conformément à leur taille (Paris / 100 000 $<$ *pop* / 20 000 $<$ *pop* $<$ 100 000 / *pop* $<$ 20 000)
- Dans chaque strate h, on détermine un nombre d'agglomérations n_h sur lesquelles l'enquête va être réalisée :

$$n_h = \mathcal{N} \times W_h$$

où $\mathcal{N}$ est le nombre total d'agglomérations enquêtées et W_h est le poids de la strate h ($W_h = \sum_{a \in h} w_a$).

- on sélectionne un échantillon de taille n_h dans la strate h. Chaque agglomération a incluse dans h a une probabilité d'inclusion proportionnelle, dans cette strate, à w_a.
- En procédant de la sorte, on peut montrer que la variance est minimale, conditionnellement au nombre total $\mathcal{N}$ d'agglomérations

Quelques mots sur l'estimateur $\hat{I}_v$ de I_v

$$\underbrace{\hat{I}_v = \sum_{a \in \mathscr{A}} \omega_a I_{a,v} \quad \text{estime sans biais} \quad I_v = \sum_{a \in A} w_a I_{a,v}}$$

$$\Updownarrow$$

$$\omega_a = \frac{w_a}{\pi_a} \text{ et } \sum_{a \in \mathscr{A}} \pi_a = \mathcal{N}$$

où $\pi_a = \Pr(a \in \mathscr{A})$ est la probabilité d'inclusion de a dans l'échantillon $\mathscr{A}$. Ce π_a est directement lié au plan de sondage de $\mathscr{A}$.

Comment sélectionner les agglos dans une strate (1) ?

Sur la probabilité d'inclusion

Comme vu précédemment, on doit adopter un plan de sondage tel que π_a est proportionnelle à w_a correspondant au poids de a dans l'indice vrai I.

Le poids w_a

correspond théoriquement au poids de a dans la dépense de consommation des ménages. On peut approximer ce poids par la démographie (lieu de résidence des ménages). Cependant, on peut faire mieux : utiliser l'enquête Budget des Familles à partir de laquelle on peut calculer le poids de a (ou du type de ville auquel appartient a) conformément au lieu d'achat.

Comment sélectionner les agglos dans une strate (2) ?

Si on ne dispose pas de contrainte d'échantillonnage

On peut sélectionner un échantillon par tirage systématique de taille n_h pour la strate h.

Mais s'il existe déjà un échantillon, on ne veut pas tout bouleverser...

Idée : on sélectionne un échantillon avec une probabilité totale d'inclusion des agglomérations proportionnelle à w_a et avec une probabilité conditionnelle d'inclusion maximale pour les agglomérations qui sont dans l'échantillon d'origine. C'est l'option retenue pour la base 2015 de l'IPC.

Comment sélectionner les agglos dans une strate (3) ?

On note $\mathscr{A}$ le nouvel échantillon d'agglomérations et $\mathscr{Y}$ l'ancien. Soit $\pi_a^{\mathscr{Y}}$ la probabilité d'inclusion de a dans $\mathscr{Y}$. On cherche $\Pr(a \in \mathscr{A} | a \in \mathscr{Y})$ tel que (1) elle est maximale et (2) la probabilité totale d'inclusion de a dans $\mathscr{A}$ est un nombre donné $\pi_a^{\mathscr{A}}$. On montre que :

- si $\pi_a^{\mathscr{A}} \leqslant \pi_a^{\mathscr{Y}}$, alors :

$$\begin{cases} \Pr(a \in \mathscr{A} | a \in \mathscr{Y}) &= \pi_a^{\mathscr{A}} \Big/ \pi_a^{\mathscr{Y}} \\ \Pr(a \in \mathscr{A} | a \notin \mathscr{Y}) &= 0 \end{cases}$$

- Si $\pi_a^{\mathscr{A}} > \pi_a^{\mathscr{Y}}$, alors :

$$\begin{cases} \Pr(a \in \mathscr{A} | a \in \mathscr{Y}) &= 1 \\ \Pr(a \in \mathscr{A} | a \notin \mathscr{Y}) &= (\pi_a^{\mathscr{A}} - \pi_a^{\mathscr{Y}}) \Big/ (1 - \pi_a^{\mathscr{Y}}) \end{cases}$$

À l'aide de ce plan de sondage conditionnel. . .

on maximise la probabilité qu'une agglomération incluse dans l'ancien échantillon $\mathscr{Y}$ soit retenue dans le nouvel échantillon $\mathscr{A}$.

Plan

Les questions complémentaires à traiter

Questions

- Quel est le "bon" nombre d'observations par variété (n_v) ?
- Quel est le "bon" nombre d'observations pour une agglomération a et pour une variété v données ($n_{a,v}$) ?

Réponses

Hypothèses : ce que nous savons

- le temps qu'un enquêteur prend pour faire une observation (dépend du type de produit)
- la variance d'observation σ_v^2 pour chaque variété

On en déduit n_v et $n_{a,v}$ par un calcul

- du minimisation de variance
- sous la contrainte que le coût de collecte total (mesuré en temps passé pour la collecte) est le même que l'actuel

Réponses en formules

- le nombre de relevés par varagglo (a, v) :

$$n_{a,v}^{\mathscr{A}} = \frac{w_a/\pi_a}{\sum_{a \in \mathscr{A}} w_a/\pi_a} \times n_v$$

- le nombre de relevé par variété (v) :

$$n_v^{\mathscr{A}} = \frac{C}{c_v} \times \frac{\kappa_v^{\mathscr{A}} \sqrt{c_v}}{\sum_v \kappa_v^{\mathscr{A}} \sqrt{c_v}}$$

où

$$\kappa_v^{\mathscr{A}} = \sigma_v w_v \sum_{a \in \mathscr{A}} w_a/\pi_a$$

C le coût total de collecte ; c_v : coût unitaire de collecte pour v ; σ_v : écart-type d'observation.

Plan

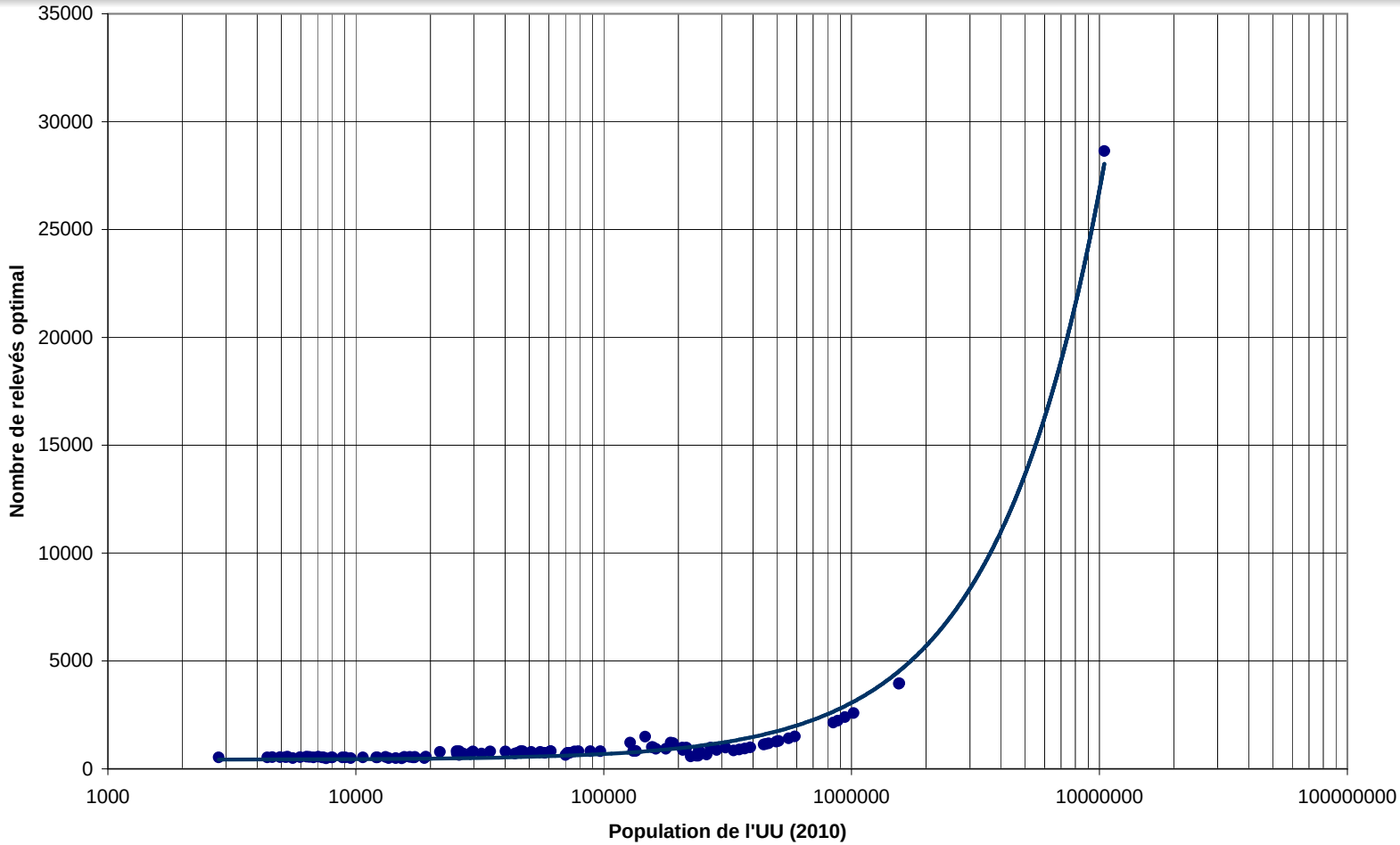
Poids des strates géographiques

Strate géographique	poids (en %)			
	démographie (1)	Alimentaire (2)	(1) normalisé sans rural	(2) normalisé sans rural
A	16.7	18.1	21.5	19.5
B2	24.6	25.6	31.7	27.6
B1	5.3	7	6.8	7.6
C	13.6	18.6	17.5	20.1
D	17.4	23.4	22.4	25.2
Rural	22.5	7.4		
Total	100	100	100	100

Durée élémentaire de collecte d'une observation de prix

type	durée sans déplacement	normalisée avec déplacement	Nombre de PV par obs.
Biens durables	102s	1.30	0.27
Habillement	60s	0.86	0.19
Alimentaire	49s	0.53	0.09
Biens manuf.	78s	1.16	0.27
Services	25s	1.73	0.61

Lien entre le nombre optimal de relevés par agglomération et la taille de l'agglomération



Nombre optimal de relevés de prix selon le type de produits

Secteur	Base 1998	Base 1998 (conservées)	Base 2015
Alimentaire	35 568	31 801	49 380
Biens durables	6 544	6 074	5 652
Habillement	21 033	20 451	11 449
Biens manuf.	29 939	27 683	29 019
Services	19 683	18 119	21 375
Total	112 767	104 128	116 875

Note : "conservées" indique les agglomérations de la base 1998 incluses base 2015.

Plan

Conclusion

Il est possible...

- de fixer les paramètres essentiels de l'échantillon pour tout couple (a, v) (i.e. $n_{a,v}$) sous réserve de disposer d'un "coût" par observation
- d'obtenir une évaluation de la variance de l'IPC pour la partie fondée sur des observations terrain (0.1 point de pourcentage (σ) sur l'évolution annuelle actuelle ; possibilité d'atteindre 0.02 avec un échantillon optimisé à moyens constants)