Optimisation de l'échantillon pour le calcul de l'indice de prix à la consommation

Pascal ARDILLY et Francis GUGLIELMETTI
Insee

ABSTRACT : Chaque indice des prix, au delà du cadre théorique général qu'imposent les utilisations habituelles de l'instrument, présente des caractéristiques spécifiques liées tant à des choix méthodologiques qu'à l'organisation de la production statistique, ce qui rend les expériences difficilement comparables.

Plusieurs pays ont précédé la France dans l'étude de la précision de leur indice (Pays-Bas, Suède, Grande Bretagne, USA par exemple). L'INSEE, dans le cadre de la rénovation de son indice des prix (qui doit aboutir en 1992), a fourni des ordres de grandeur de sa précision, à chaque niveau d'agrégation. Après un exposé de la détermination des échantillons et des formules de calcul qui en découlent (1e partie), sont présentés ici les résultats les plus importants concernant la précision (2e partie) puis, prolongement naturel de ce travail, la répartition des échantillons assurant en théorie la meilleure précision possible (3e partie).

KEY WORDS : price index/variance estimation/sample optimization

1. EXPRESSION ET CALCUL DE L'INDICE DES PRIX

L'indice des prix calculé en France par l'INSEE est défini comme un "indice des prix à la consommation des ménages urbains dont le chef est ouvrier ou employé". Le champ de l'indice est donc l'ensemble des biens et services consommés par les ménages de référence.

On détermine en premier lieu une nomenclature des produits, qui constitue une stratification du champ, chaque strate étant appelée poste de dépenses. Les postes sont regroupés en secteurs, au nombre de quatre : Alimentation, Habillement, Autres produits manufacturés et Services. Un poste rassemble des produits en très grand nombre, dont la liste varie d'ailleurs chaque jour. On décide de le suivre au travers de quelques représentants "bien choisis" de ces produits, appelés variétés. C'est au sein de chaque variété que l'on tire les relevés de prix qui constitueront les unités d'observation. On distingue deux types de variétés : la variété homogène, composée de produits dont les prix sont directement comparables (la baguette de pain par exemple), et la variété hétérogène, regroupant des produits de même fonction, mais dont on ne peut pas ajouter les prix (la poupée d'enfant par exemple).

L'indice vrai, inconnu et estimé, est un indice de

prix de type Laspeyres défini entre deux dates données. Il met en jeu des transactions réelles.

La grandeur de base utilisée dans l'expression de l'indice est le prix de la transaction élémentaire (j) dans l'agglomération (i) à la date (t) pour chaque variété (v) retenue, soit $p^t(j,i,v)$

. Notons (t_o) la date de référence, et N(i,v) le nombre de relevés potentiels dans l'agglomération (i) pour la variété (v). L'indice élémentaire agglomération-variété est :

$$I(i,v) = \frac{\overline{p}^t(i,v)}{\overline{p}^{t_o}(i,v)} \quad \text{où} \quad \overline{p}^t(i,v) = \frac{1}{N(i,v)} \sum_{j=1}^{N(i,v)} p^t(j,i,v)$$

si la variété est homogène, et :

$$I(i,v) = \frac{1}{N(i,v)} \sum_{j=1}^{N(i,v)} \frac{p^t(j,i,v)}{p^{t_o}(j,i,v)}$$

si la variété est hétérogène.

Ces indices élémentaires sont à la base des expressions de tous les indices synthétiques, pour lesquels on peut distinguer deux "dimensions" : la dimension géographique et le niveau d'agrégation des produits.

On définit 5 catégories de communes (notées CC) :

CC2 : Agglomérations de 2 000 à 10 000 habitants

4 : " 10 000 à 100 000 "

6 : " 100 000 à 200 000 "

8 : " de plus de 200 000 "

9 : Agglomération de Paris.

Les CC sont croisées avec des groupements de régions administratives appelées ZEAT, au nombre de 8 (Région Parisienne, Nord, Ouest, etc...). On obtient des indices aux niveaux CC, ZEAT, France entière, pour la variété, le poste ou l'ensemble des produits par pondération des indices élémentaires I(i,v). Ainsi, l'indice de variété France entière est :

$$I(v) = \sum \frac{C(i,v)}{C(v)} I(i,v) \text{ , avec } C(v) = \sum_{i=1}^{M} C(i,v)$$

où C(i,v) est la consommation de la variété (v) dans l'agglomération (i), et M le nombre total d'agglomérations de plus de 2 000 habitants. De même, l'indice de poste France entière vaut :

$$I(p) = \sum_{v \in p} w(v) \cdot I(v) \text{ , avec } \sum_{v \in p} \omega(v) = 1 \text{ ,}$$

où w(v) est le poids économique de la variété (v) dans le poste (p). L'indice global actuel comprend 296 postes et un peu plus de 1 000 variétés.

La détermination de l'échantillon de prix résulte d'un sondage à deux degrés : tirage d'agglomérations, puis tirage de relevés au sein de ces agglomérations.

Le tirage des unités primaires s'effectue de la manière suivante : considérant l'ensemble des agglomérations de plus de 2 000 habitants, on réalise une stratification préalable par CC-ZEAT. On retient d'office toutes les agglomérations

de CC8, ainsi que Paris. Pour les trois autres CC, on subdivise chaque strate en "groupes" selon un critère de taille. Dans un premier temps, on réalise un tirage systématique sur les agglomérations de la strate classées par groupe, selon un pas fonction de la CC : en CC6, par exemple, on retient une agglomération pour 250 000 habitants. On ne s'intéresse, dans cette opération, qu'au nombre d'agglomérations retenues dans le groupe (g), soit m_g. Dans un deuxième temps, on tire par sondage aléatoire simple dans chaque groupe, un nombre d'agglomérations égal à m_g. Cette méthode en deux temps se justifie par le souci de conserver au maximum l'échantillon d'agglomérations résultant du précédent tirage. Le tirage aléatoire simple utilisé ne sert en fait qu'à compléter l'échantillon pré-existant si le groupe a été désigné trop souvent par rapport à l'ancien tirage, ou à supprimer le plus aléatoirement possible des agglomérations dans le cas contraire. La taille de l'échantillon par groupe est aléatoire, et la probabilité d'inclusion d'une agglomération est à peu près proportionnelle à la taille moyenne des agglomérations du groupe auquel elle appartient. Le dernier tirage d'agglomération a été réalisé pour l'indice base 100 en 1970, avec l'information disponible à l'époque. Le prochain tirage sera fait pour l'indice rénové en 1992.

Dans une agglomération de l'échantillon pour une variété donnée, on considère que le tirage des relevés s'effectue

par sondage aléatoire simple. Pour des raisons évidentes de commodité, on ne relève pas les prix des transactions mais les prix affichés. Cela revient à faire l'hypothèse que toutes les transactions concernant chacun des produits observés sur la période considérée sont réalisées au même prix. Cette période est le plus souvent le mois, qui est la période de calcul de l'indice.

Si on note n(i,v) le nombre de relevés effectués dans l'agglomération (i) pour la variété (v), alors les estimateurs des véritables indices sont :

$$\hat{I}(i,v) = \frac{\bar{p}^{t}(i,v)}{\bar{p}^{t_0}(i,v)} \quad , \text{ où } \quad \bar{p}^{t}(i,v) = \frac{1}{n(i,v)} \times \sum_{j=1}^{n(i,v)} p^{t}(j,i,v)$$

si la variété est homogène, et :

$$\hat{I}(i,v) = \frac{1}{n(i,v)} \times \sum_{j=1}^{n(i,v)} \frac{p^{t}(j,i,v)}{p^{t_0}(j,i,v)}$$

si la variété est hétérogène.

La taille n(i,v) est actuellement déterminée en fonction d'estimations de la consommation de la variété dans la CC, et d'une idée a priori de la dispersion des prix dans la CC. Cependant, elle obéit toujours aux contraintes de terrain et à des grilles de répartition anciennes et empiriques ; c'est précisément l'objet de l'optimisation de tenter d'améliorer les allocations par CC-variété.

Le but des calculs de variance menés ici est de fournir les ordres de grandeur des précisions des indices publiés, ordres de grandeur que nous ne connaissions pas auparavant. Ces ordres de grandeur sont d'autant plus fiables que le niveau d'agrégation est grand. L'important est de savoir, au moins pour les indices de secteur et l'indice national, si les valeurs publiées reflètent, disons à quelques dizièmes de points, la réalité, ou si, au contraire, l'inflation est mal traduite par ces indices.

Si les estimations sont à manipuler avec prudence, c'est parce que (et c'est d'ailleurs inévitable si on veut conserver un budget raisonnable) un grand nombre de simplifications et d'hypothèses sont nécessaires pour parvenir à la situation idéale que traduit le plan de sondage simple précédemment décrit. Avant de passer aux résultats, voyons les principales difficultés rencontrées.

Nous manipulons des échantillons de petite, voire très petite taille. Les calculs élémentaires se font au niveau variété-CC, et si l'échantillon global comporte environ 130 000 relevés (hors produits frais), ceux-ci se répartissent sur 105 agglomérations (initialement 108) pour 1 000 variétés, ce qui conduit théoriquement à moins de deux relevés en moyenne par agglomération ! On ne peut donc pas attendre une grande stabilité des estimateurs de variance à un niveau fin.

Le tirage initial des agglomérations ne pose pas de problème : on peut considérer que l'aspect aléatoire est bien respecté. Par contre la désignation des variétés résulte plus souvent d'un choix raisonné que d'un tirage au sort : les critères de sélection sont la part de la consommation du poste attribuée à la variété, le fait que la variété puisse avoir un comportement, en matière d'évolution de prix, voisin de l'ensemble des produits du poste (d'après dires d'experts), et enfin la facilité de suivi des produits composant cette variété. En conséquence, on ne calculera pas de précision liée au tirage des variétés, considérant que l'erreur due à cette opération est plutôt de la nature du biais. Le nombre important de variétés, la présence de variétés "Autres" dans les postes très hétérogènes, les avis de professionnels sur la comparaison des évolutions de prix des produits du poste, sont des éléments qui concourent à minimiser le biais de l'indice. Par contre, l'inexactitude des poids attribués aux variétés en question peut être néfaste : en effet, si le poids du poste pour chaque CC est connu par les enquêtes Budget et, chaque année, par la Comptabilité Nationale, il faut utiliser une grille de répartition plus ou moins approximative pour déterminer la part de consommation due à la variété.

Enfin, le tirage des relevés comporte des aspects délicats : on laisse les enquêteurs déterminer eux-mêmes les

produits à suivre dans la variété, puisqu'on ne dispose pas de base de sondage au niveau produit. On leur fournit le nombre total de relevés, ainsi que des consignes de "représentativité" selon les grands types de points de vente (petits commerçants, supermarchés,...) et selon certains groupes de produits que l'on peut discerner dans les variétés hétérogènes. L'enquêteur est tenté, pour minimiser ses déplacements, d'effectuer le plus de relevés possible dans un même point de vente (sous la seule contrainte de ne pas y faire deux relevés pour la même variété). Il introduit ainsi un effet de grappe en tirant des points de vente au lieu de tirer des relevés.

2 . PRECISION

2.1. INDICE DE VARIETE

La première difficulté rencontrée dans le calcul de la variance de l'indice CC-variété est liée à la détermination des poids des agglomérations pour la variété considérée. Les poids CC-poste sont connus de façon précise, mais les poids de la variété dans la CC sont le plus souvent estimés. Ils doivent être, dans une ultime étape, répartis entre les agglomérations de l'échantillon initial. Malheureusement, certaines variétés ont des poids trop faibles pour pouvoir être représentées dans toutes les agglomérations tirées. Par ailleurs, il y a des agglomérations où la variété n'existe pas. Si on tient compte, en sus, des inévitables problèmes de terrain, on obtient finalement un tableau de pondérations (et par suite de relevés) croisant variétés et agglomérations qui comporte beaucoup de cases vides. A dire vrai, et c'est là la difficulté, les règles d'affectation demeurent empiriques, la seule contrainte à peu près respectée étant de limiter la dispersion des poids entre les agglomérations. Encore trouve-t-on des CC où la variété n'est pas du tout représentée ! La présence de ces trous amène à utiliser, dans les calculs d'indice, un estimateur de la forme :

$$\hat{I}(cc,v) = \sum_{i \in cc} w(i,v,s) \times \hat{I}(i,v)$$

où le poids w(i,v,s), attaché à l'agglomération (i) pour la variété (v) est une fonction complexe de l'échantillon (s) d'agglomérations, et même, plus exactement, de l'échantillon d'agglomérations-variétés.

L'estimateur utilisé dans les calculs de variance attribue un poids identique à chaque agglomération. Il s'agit de :

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

où m(cc,v) est le nombre total d'agglomérations dans la CC où la variété (v) est relevée. L'égalité des poids se justifie si on considère que, dans la CC, le tirage des agglomérations est effectué proportionnellement à la taille moyenne de l'agglomération du groupe (probabilité d'inclusion $\pi(i)$)

L'indice vrai étant :

$$I(cc,v) = \sum_{i \in cc} w(i,v) \times I(i,v)$$

l'estimateur sans biais vaut :

$$\hat{I}(cc,v) = \sum_{i \in s} \frac{w(i,v)}{\pi(i)} \times \hat{I}(i,v)$$

Comme w(i,v) et $\pi(i)$ sont, à priori, à peu près proportionnels à la taille de l'agglomération (i), on obtient finalement un poids constant.

Lorsque la variété est hétérogène, l'estimateur est sans biais car il a la forme d'une moyenne simple. Lorsque la

variété est homogène, il s'agit d'un estimateur classique du type ratio : le biais peut alors être important au niveau d'une agglomération, voire pour une CC, si la variété est peu consommée. L'indice national de la variété, par contre, doit avoir un biais négligeable (une variété moyenne comporte 120 relevés).

Nous n'avons pas considéré le système de tirage de groupe, car, pour certaines variétés, l'échantillon d'agglomérations enquêtées est trop faible : plutôt que de manipuler un tirage stratifié avec taille aléatoire dans chaque strate (c'est-à-dire en pratique une taille valant 0 ou 1, exceptionnellement 2), nous avons utilisé le schéma d'un tirage aléatoire simple dans la CC, où, pour les CC2, 4 et 6, la taille d'échantillon était fixée et égale à celle qui avait été empiriquement déterminée lors de l'affectation des poids. Pour la CC8, au contraire, on considère que les relevés doivent être effectués dans toutes les agglomérations, et on met au compte de l'"accident" les cas où ce n'est pas vérifié.

Dans ces conditions, la variance de l'indice d'une variété homogène prend la forme classique :

$$V\left[\hat{I}(cc,v)\right] = V_1(cc,v) + V_2^{HOM}(cc,v)$$

$$V_1(cc,v) = \frac{1}{m(cc,v)} \times \left(1 - \frac{m(cc,v)}{M(cc)}\right) \times S^2(cc,v)$$

$$V_2^{HOM}(cc,v) = \frac{1}{m(cc,v) \times M(cc)} \times \sum_{i=1}^{M(cc)} \frac{S^2(HOM,i,v)}{m(i,v)}$$

avec :

$$S^2\ (CC,V) = \frac{1}{M(CC)\ -\ 1} \sum_{i=1}^{M(CC)} \left(I(i,v) - \frac{1}{M(CC)} \sum_{k=1}^{M(CC)} I\ (k,v) \right)$$

$$S^2\,(HOM,i,v) = \frac{1}{\left(\overline{P}^{t}o\,(i,v)\right)^2} \times \frac{1}{N(i,v)-1} \times \sum_{j=1}^{N(i,v)} \left(P^t\,(j,i,v)-I(i,v)\ \times\ P^to\,(i,j,v)\right)^2$$

M(cc) est le nombre d'agglomérations dans la cc, et N(i,v) est le nombre de relevés potentiels dans l'agglomération (i) pour la variété (v).

V_1 et V_2 représentent respectivement les termes inter-agglomérations et intra-agglomération du tirage à deux degrés.

Si la variété est hétérogène, alors :

$$V\ [\hat{I}\ (CC,C)] = V_1\ (CC,V) + V_2^{HET}\ (CC,\ V)$$

$$V_2^{HET}\ (CC,V) = \frac{1}{m\ (CC,V)\ \times\ M\ (CC)} \times \sum_{i=1}^{M\ (CC)} \frac{S^2\,(HET,i,v)}{n\ (i,v)}$$

avec : $$s^2\ (HET,i,v) = \frac{1}{N\ (i,v) - 1} \times \sum_{j=1}^{N(i,v)} \left(\frac{p^t\ (j,i,v)}{p^{to}\ (j,i,v)} - I\ (i,v) \right)^2$$

Les estimateurs sans biais de V_1, et V_2^{HOM} sont respectivement :

$$\hat{V}_1\ (CC,V) = \frac{1}{m\ (CC,v)} \times \left(1 - \frac{m\ (CC,V)}{M\ (CC)}\right)$$

$$\times \left(\frac{1}{m\ (CC,V) - 1} \times \sum_{j=1}^{m\ (CC,V)} (\hat{I}(\ ,v) - \hat{I}\ (CC,V) \right)^2$$

$$- \frac{1}{m(CC,V)} \times \sum_{i=1}^{m(CC,V)} \frac{s^2\ (HOM,\ i,v)}{n\ (i,v)} \Bigg)$$

$$\hat{V}_1\ (CC,V) = \frac{1}{(m\ (CC,V))^2} \sum_{i=1}^{m(CC,V)} \frac{s^2\ (HOM,\ i,\ V)}{n\ (i,v)}\ ,$$

où $$s^2\ (HOM,\ i,v) = \frac{1}{(\bar{P}^{to}\ (i,v))^2} \times \frac{1}{n\ (i,v) - 1}$$

$$\sum_{j=1}^{n(i,v)} (p^t\ (j,i,v) - \hat{I}\ (i,v) \times p^{to}\ (j,i,v))^2$$

Si la variété est hétérogène, il suffit de remplacer l'estimateur $s^2\ (HOM,i,v)$ par l'expression :

$$s^2\ (HET,i,v) = \frac{1}{n(i,v) - 1} \times \sum_{j=1}^{n(i,v)} \left(\frac{p^t\ (j,i,v)}{p^{to}\ (j,i,v)} - \hat{I}\ (i,v) \right)^2$$

On notera que la variance inter-agglomération ne concerne que les CC2, 4 et 6, classes d'agglomérations représentant à peu près la moitié de la consommation nationale.

L'ancienneté du tirage, et donc de la base de sondage, rend les estimateurs en théorie biaisés. En effet, certaines agglomérations sont comptabilisées dans une strate qui n'est plus actuellement la leur. Si on considère que les mouvements de prix sont liés à la CC, les indices estimés ne sont plus sans biais du véritable indice de la CC. Replacer les agglomérations dans leur véritable CC n'arrange rien puisqu'on estime les dispersions avec une formule de sondage aléatoire

simple alors que les probabilités de tirage sont inégales et fonction de la CC d'origine.

En toute rigueur, il faudrait considérer l'aléa lié au tirage stratifié d'origine mais calculer la précision d'un indice post-stratifié avec les strates actuelles. Considérant que l'effet CC explique peu le comportement de l'indice, que les changements de strate sont marginaux et que les estimations et la répartition des poids ne tiennent pas toujours correctement compte des modifications de taille des agglomérations (ce qui rendrait illusoire les améliorations théoriques envisageables), nous avons choisi de laisser les agglomérations dans leur strate d'origine.

Le cas le plus favorable au calcul est celui où la variété est présente dans au moins deux agglomérations de chacune des trois CC : 2, 4 et 6. Dans ce cas, l'écart-type inter-agglomérations final de cette variété est :

$$\hat{\sigma}(v, \text{INTER}) = \sqrt{\sum_{cc=2,4,6} w^2(cc|v) \times \hat{V}_1(cc, v)}$$

avec $w(2|v) + w(4|v) + w(6|v) + w(8|v) + w(9|v) = 1$

Si la variété est présente dans au moins une agglomération de chaque CC (elle a donc un poids dans chaque CC), mais que, dans une ou deux de ces CC, elle n'est relevée en fait que dans une seule agglomération, alors, par une règle de proportionnalité, on estime la dispersion $S^2(cc,v)$ par :

$$\hat{S}^2(cc,v) = \frac{S^2(cc)}{S^2(cc^*)} \times S^2(cc^*,v)$$

où CC* est la CC où un calcul de dispersion est possible, et $S^2(cc)$ est la dispersion des indices d'agglomération de la CC calculée surl'ensemble des variétés. Si par contre, il existe une CC où la variété n'est pas du tout relevée, alors on décide de laisser à blanc la variance totale. En effet, le poids de cette variété est (artificiellement) nul, et il est impossible de le modifier sans que l'ensemble de la structure des poids le soit. On considère que cette situation est anormale, temporaire, et entièrement due aux impératifs de terrain ; il n'est pas légitime de prendre en compte une variance nulle : si l'échantillon, n'étaitpas de taille nulle, il serait petit, et donc donnerait lieu à une variance plutôt grande. On laisse donc la variance finale à blanc, préférant reporter sa prise en compte au niveau du poste.

En ce qui concerne l'estimation de la variance intra-agglomération, on peut faire les remarques suivantes : la

pratique de l'INSEE est d'effectuer, autant que possible, au moins deux relevés dans chaque agglomération où l'on décide de faire des relevés, ce qui permet, en théorie, d'estimer une dispersion. En réalité, disposer de deux relevés à chaque date ne signifie pas nécessairement avoir deux relevés suivis entre les deux dates, et des valeurs manquantes peuvent survenir à cette occasion. Si le produit a été remplacé plusieurs fois, on reconstitue mal la chronique de son indice. Chaque fois que cela a été possible, tous les remplacements ont été assimilés à des remplacements équivalents ; cela est sûrement abusif, mais les cas où le produit remplaçant ne peut être comparé au produit remplacé sont relativement rares (moins de 1 % des produits) et cette procédure va plutôt dans le "bon sens" (augmentation artificielle de la dispersion).

Lorsque le poids d'une variété-agglomération est non nul, mais que la dispersion n'est pas calculable, on estime celle-ci par la moyenne des dispersions de la variété dans les autres CC (ou par la moyenne des dispersions des autres variétés dans la même CC, ce qui ne modifie pas les valeurs finales). Si la variance intra-agglomération est calculable dans chaque CC, sauf éventuellement Paris, on calcule :

$$\hat{\sigma}(v, \text{INTRA}) = \sqrt{\sum_{cc=2}^{9} w^2(cc|v) \times \hat{V}_2(cc, v)}$$

Sinon, dès qu'il existe une CC différente de la CC9 où l'estimation est impossible, on laisse l'écart-type à blanc. S'il n'y a aucun relevé à Paris, on considère que la consommation y est réellement négligeable, et que la variance correspondante est nulle (peu de variétés sont ici concernées).

Enfin, on introduit un traitement spécifique à certaines variétés dont les prix sont fixés à certains niveaux géographiques : on distingue les tarifs nationaux, où le prix est unique (tabacs, téléphone, timbres, etc...), et les tarifs locaux, où le prix est fonction seulement de l'agglomération. Pour les premiers, la variance totale est nulle ; pour les seconds, seule la variance intra est nulle.

2.2. INDICE DE POSTE ET INDICES AGREGES

L'indice de poste est une moyenne pondérée des indices de variété. Sa précision comprend un terme de covariance, et a pour expression :

$$V(\hat{I}(CC,P)) = \sum_{V \in P} W_{cc}^2(V/P) \times V(\hat{I}(cc,v)) + \sum_{V \neq V'} \frac{W_{cc}(V/P) \times W_{cc}(V'/P)}{m(CC,V) \times M(CC)}$$

$$\times \left(1 - \frac{m(CC,V)}{M(CC)}\right) \times \frac{1}{M(CC) - 1} \times \sum_{i=1}^{M(CC)} (I(i,v) - \bar{I}(v)) \times (I(i,v') - \bar{I}(v'))$$

avec : $w_{cc}(v|p) = \frac{w(v,cc)}{w(p,cc)}$, poids de la variété dans le poste au sein de la CC, et :

$$\bar{I}(v) = \frac{1}{M(cc)} \times \sum_{i=1}^{M(cc)} I(i,v)$$

Il n'apparaît pas de terme intra dans la covariance parce qu'on peut supposer, en première approximation, que les tirages de relevés sont indépendants. Pourtant, l'enquêteur peut tirer en réalité des points de vente, et un point de vente plus cher que la moyenne pour une variété a de grandes chances, a priori, d'être plus cher que la moyenne pour une autre variété du même poste. On négligera cet éventuel effet de grappe. Le terme de covariance lié au premier degré est difficile à estimer, car il faut que, pour chaque couple de variétés, on puisse trouver au moins deux agglomérations dans la CC où les deux variétés du couple soient relevées simultanément. Sur quelques postes représentatifs où le calcul était possible, on a constaté que, d'une part il existait des covariances positives et négatives qui se compensaient en partie, et, d'autre part, sauf pour les postes ayant de nombreuses variétés, le terme de cova-

riance était petit devant le terme de variance. On obtient finalement la précision de l'indice de poste France entière selon :

$$\hat{V}\left[\hat{I}(\text{FRANCE, POSTE})\right] = \sum_{cc} w^2(cc|p) \times \hat{V}\left[\hat{I}(cc, p)\right]$$

$$= \sum_{v \in p} w^2(v|p) \times \hat{V}\left[\hat{I}(\text{FRANCE}, v)\right]$$

avec : $w[cc|p] = \frac{w(cc,p)}{w(p)}$, poids de la catégorie de commune dans le poste et

$$\hat{V}\left[\hat{I}(\text{FRANCE}, v)\right] = \sum_{cc} w^2(cc|v) \times \hat{V}\left[\hat{I}(cc, v)\right]$$

Lorsqu'une au moins des variétés composant le poste a une précision à valeur manquante, on effectue deux traitements parallèles : soit on repondère les variétés renseignées du poste de façon à respecter l'égalité des poids à un, soit on affecte aux variétés non renseignées la précision moyenne des variétés renseignées du poste, et on pondère la variété par son véritable poids. C'est donc à ce niveau qu'on prend en compte l'aléa existant dans la ou les CC où l'estimation n'avait pu être réalisée au niveau variété. Le premier traitement fournit des écarts-types plus grands que le second. Pour quelques postes où aucune précision de variété n'est calculable, on laisse l'écart-type final du poste à blanc.

Il suffit, pour obtenir des précisions d'indices plus agrégés, de pondérer les variances des indices de poste par les

carrés des poids adéquats. On considère en effet qu'au delà du niveau poste, la covariance ne joue plus.

On peut passer du poste au secteur de la même façon que l'on est passé de la variété au poste, c'est-à-dire soit par repondération, soit par imputation sur les dispersions manquantes. On obtient ainsi quatre résultats par secteur selon quatre scénarios. Ceux-ci se traduisent par quatre précisions proposées au niveau ensemble des produits.

2.3. RESULTATS

Les calculs de précision des indices nécessitent la mobilisation de l'ensemble de l'information élémentaire (relevés de prix) à deux dates. Sauf indications contraires, les résultats concernent les années 1987 et 1988 et portent sur l'indice annuel, plus précisément sur la variation des prix entre deux mois de décembre consécutifs.

Ce choix s'explique pour plusieurs raisons : l'importance accordée par le public à la variation sur une année, la difficulté d'interpréter les variations mensuelles (en particulier en raison de la périodicité trimestrielle de certains relevés), la mise à jour annuelle de l'échantillon de relevés. Rappelons que l'indice français est un indice-chaîne ; son rebasage chaque mois de décembre facilite la mise à jour de l'échantillon de relevés mais, par suite, rend plus difficile le suivi de produits identiques sur une période pluri-annuelle.

Les précisions sont données par les estimations des écarts-types des indices estimés. L'unité de compte est le point d'indice et non un pourcentage d'évolution. Ces estimations permettent de proposer un intervalle de confiance pour la vraie valeur des indices concernés : l'indice du pain ayant augmenté de 3,8 % entre décembre 1986 et décembre 1987, son écart-type étant 0,1, il y a environ 95 chances sur 100

pour que l'indice vrai du pain sur la période soit compris entre 103,6 et 104,0.

Les calculs de précision des indices de poste et de variété appellent les remarques suivantes (tableau 1 ; graphiques 1 et 2) :

- la précision des indices de variété est mauvaise, voire très mauvaise. Cela n'est guère étonnant si l'on pense qu'elle dépend essentiellement du nombre de relevés ; ce nombre étant largement déterminé par la pondération, chaque fois que celle-ci est faible (disons quelques unités sur 10 000), le nombre de relevés ne dépasse pas la centaine, d'où une grande imprécision de l'indice estimé. Les variétés dont les écarts-types calculés sont supérieurs à 2 représentent 11 % des variétés mais seulement 2 % de la pondération. L'écart-type médian est 1,1.

- l'écart-type inter-agglomérations est généralement faible par rapport à l'écart-type intra-agglomération : environ 35 % des variétés ont un estimateur de l'écart-type inter négatif (ce qui laisse supposer que le véritable écart-type inter est très faible). Pour les variétés où l'estimateur est positif, le rapport moyen de l'écart-type inter à l'écart-type intra est de 0,4.

- la précision des indices de poste est un peu meilleure mais reste globalement médiocre : 50 % des indices de poste ont un écart-type calculé supérieur à 0,7. Ce résultat revêt une importance beaucoup plus grande que pour les variétés, parce que les indices des postes sont publiés. Les postes pour lesquels les estimateurs sont supérieurs à 1 représentent 25 % des postes, mais seulement 8 % de la pondération ; les "mauvais" postes sont (le plus souvent) de "petits" postes.

- ces résultats sont quasiment les mêmes en 1987 et 1988 (cela explique que les courbes des graphiques 1 et 2 représentent ces deux années). Parmi les 50 postes ayant les indices estimés les moins précis, on en retrouve 30 identiques les deux années.

Aux niveaux supérieurs d'agrégation, les précisions sont naturellement meilleures (tableau 2) :

- les résultats des deux années consécutives confirment la très grande stabilité de la qualité des indices.

- les précisions sont très différentes d'un secteur à l'autre et on note à nouveau la relation qui unit la précision au nombre de relevés. Le secteur des Services a un indice plus précis que celui de l'Habillement, malgré un nombre de relevés moindre, parce qu'il "bénéficie" de la présence des tarifs, dont les indices ont une dispersion nulle.

- les quatre scénarios concernant le traitement des précisions manquantes fournissent des résultats assez voisins pour l'indice général compte tenu des hypothèses faites,plutôt différents au niveau sectoriel, en particulier pour les Services, où les valeurs manquantes représentent plus du quart de la pondération. On note, sans pouvoir bien l'expliquer, qu'en cas de précision manquante pour une variété, le remplacement par une valeur moyenne fournit une précision meilleure que la repondération pour l'indice du poste alors qu'appliquées au niveau d'un poste, les deux techniques ont l'effet inverse sur les indices agrégés.

- une précision de l'indice général estimée par un écart-type voisin de 0,05 est un résultat très satisfaisant : si on accepte d'utiliser l'expression classique des intervalles de confiance, quand l'indice calculé pour 1987 est 103,1, on peut dire, avec une probabilité proche de 0,95, que l'indice vrai se situe entre 103,0 et 103,2.

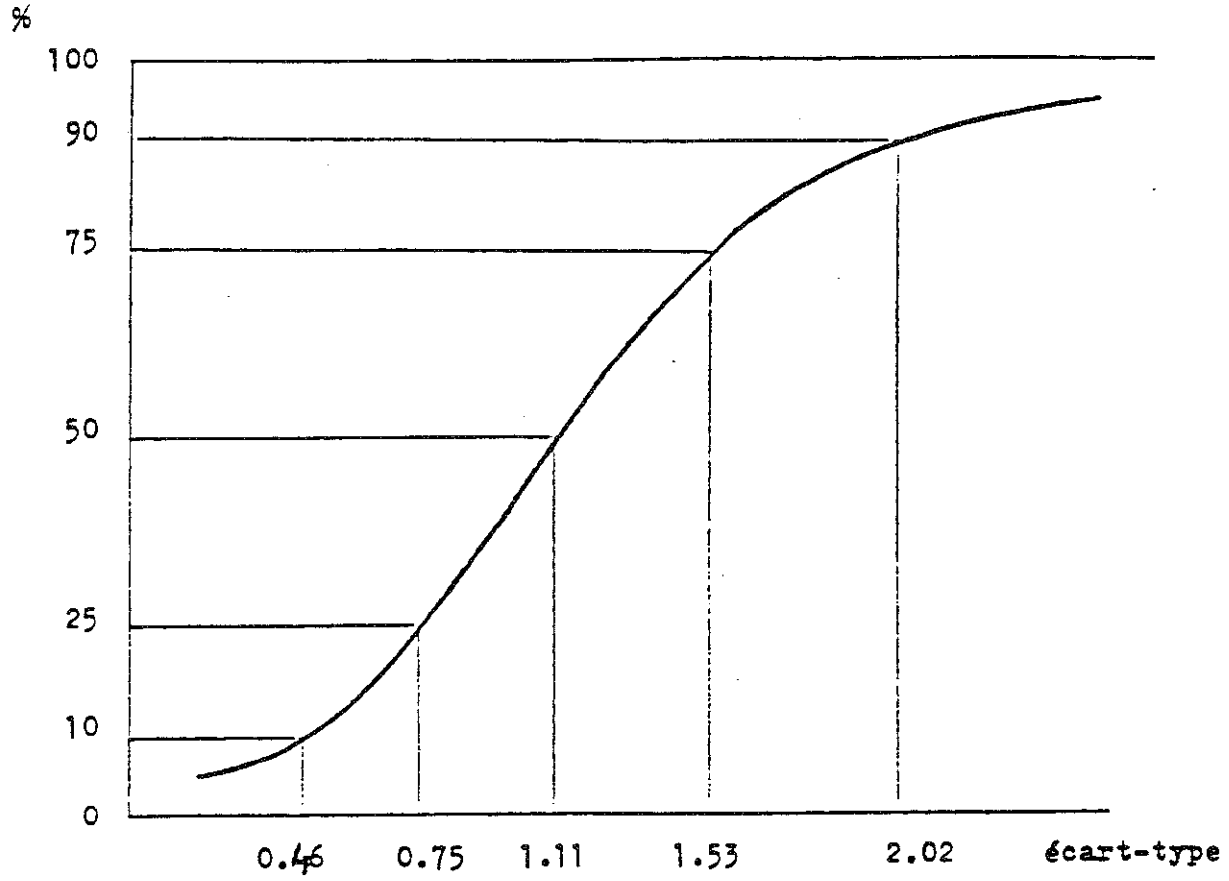
TABLEAU 1
QUELQUES EXEMPLES DE PRECISION DES INDICES DE POSTE ET DE VARIETE (1987)

Poste/Variété	Ecart type inter	Ecart type intra	Ecart type global	Pondération (1/10000e)	Nombre de relevés
PAIN			0.11	95	2 200
Pain de 400g	0.09	0.12	0.15	45	1 000
BOEUF A ROTIR			0.15	144	3 000
Entrecôte	0.20	0.38	0.43	18	560
CONSERVES DE FRUITS			1.44	3	200
Abricots au sirop	-	3.05	2.79	1	50
ROBES POUR FEMMES			0.92	50	900
Robe en tissu synthétique	-	1.11	1.08	32	600
CHEMISES POUR HOMMES			0.71	26	700
Chemise de ville 100 % coton	0.96	1.18	1.52	4	100
OUTILLAGE, QUINCAILLERIE			0.42	20	700
Vis à bois	-	1.75	1.33	1	50
Panneau d'aggloméré	0.06	0.72	0.72	4	170
ESSENCES			0.12	400	1 800
Super	0.05	0.13	0.14	333	950
ENTRETIEN COURANT DES VEHICULES			0.94	30	500
Huile pour moteur	-	0.71	0.66	10	170
BOISSONS CHAUDES AU CAFE			0.45	40	500
Tasse de café	0.05	0.54	0.54	27	270

N.B. " - " : estimateur négatif

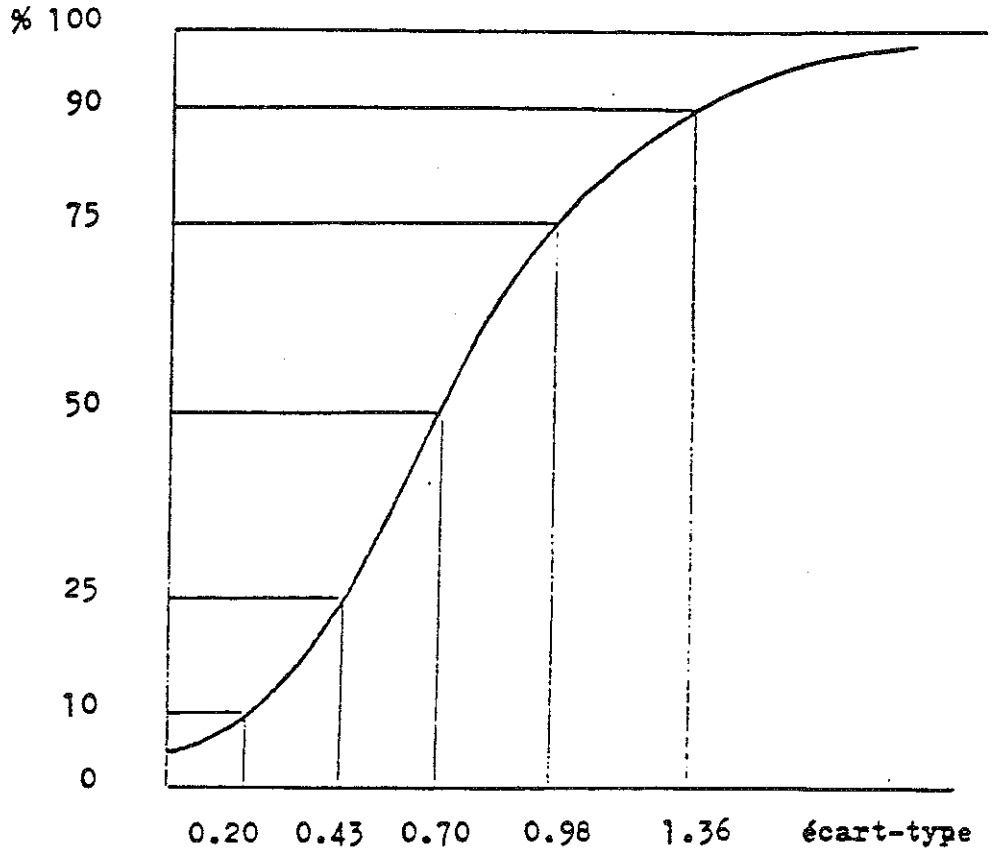
Graphique 1

REPARTITION DES INDICES DE VARIETE SUIVANT LEUR PRECISION
(% DES INDICES AYANT UN ECART-TYPE INFERIEUR A UNE CERTAINE VALEUR)



Graphique 2

REPARTITION DES INDICES DE POSTE SUIVANT LEUR PRECISION
(% DES INDICES AYANT UN ECART-TYPE INFERIEUR A UNE CERTAINE VALEUR)



N.B. La courbe correspond aux écarts-types calculés avec repondération des variétés de précision connue.

Tableau 2

PRECISION DES INDICES DE SECTEURS ET DE L'INDICE GENERAL

1987	Pondération (%)	Nbre de relevés (1 000)	(1)	(2)	(3)	(4)	Moyenne
Alimentation	22	44	0.063	0.063	0.059	0.059	0.061
Habillement	10	23	0.149	0.148	0.133	0.132	0.140
Autres Manufacturés	35	44	0.061	0.088	0.059	0.083	0.073
Services	33	19	0.100	0.114	0.073	0.098	0.096
Ensemble	100	130	0.043	0.053	0.036	0.047	0.045

1988	Pondération (%)	Nbre de relevés (1 000)	(1)	(2)	(3)	(4)	Moyenne
Alimentation	21	43	0.063	0.063	0.059	0.059	0.061
Habillement	10	23	0.148	0.148	0.133	0.133	0.140
Autres manufacturés	35	45	0.059	0.085	0.055	0.080	0.070
Services	34	19	0.083	0.104	0.056	0.086	0.082
Ensemble	100	130	0.039	0.050	0.033	0.044	0.042

Traitement des précisions marquantes :

(1) (Variété : repondération) et (poste : repondération)
(2) (Variété : repondération) et (poste : imputation de précision moyenne)
(3) (Variété : imputation de précision moyenne) et (poste : repondération)
(4) (Variété : imputation de précision moyenne) et (poste : imputation de précision moyenne)

3. OPTIMISATION

De façon à diminuer les coûts sans modifier la précision, nous avons cherché à optimiser l'échantillon d'agglomérations. Une fois la liste des agglomérations connue, il est encore possible d'optimiser la répartition des relevés par variété entre ces agglomérations. L'optimisation a donc été réalisée en deux temps.

3.1 ECHANTILLON D'AGGLOMERATIONS

Elle s'effectue sur l'indice général en une seule fois, en tenant compte de l'ensemble des variétés, et en attribuant à chacune une importance proportionnelle à son poids économique. Chaque variété étant un problème en soi, on devrait effectuer une optimisation par variété. Cependant on y renonce pour les raisons suivantes : les estimateurs obtenus seraient trop instables pour que l'on puisse raisonnablement les utiliser et faire dépendre d'eux le réseau d'enquêteurs ; on ne contrôlerait plus le nombre total d'agglomérations, qui deviendrait, pour certaines variétés, excessif, et l'instabilité dans le temps occasionnerait des problèmes considérables de gestion, alors qu'on cherche plutôt à limiter au maximum les modifications d'organisation. Enfin, cette optique est un peu contradictoire avec la recherche d'une meilleure qualité d'ensemble, sachant qu'au niveau de la variété, il est impossible, compte

tenu du budget, d'obtenir de bons résultats. Par conséquent, on résoud le programme :

$$MIN \left(\sum_{cc,z} m(cc,z) \right)$$

$$sc \left| \sum_{cc,z,v} w^2(cc,z,v) \times \left(1 - \frac{m(cc,z)}{M(cc,z)}\right) \times \frac{S^2(cc,z,v)}{m(cc,z)} = \tilde{V} \right.$$

(z) est l'indice courant de la ZEAT, et $\tilde{V}$ la variance calculée avec l'échantillon actuel.

Les dispersions sont calculables pour les années 81 à 87. On peut donc apprécier la stabilité des résultats dans le temps. La détermination des effectifs se fait au niveau CC-ZEAT, car il s'agit du découpage géographique de base, qui coïncide avec la définition des strates utilisées lors du tirage. La notion de groupe disparaît, l'effectif d'agglomérations dans certaines CC-ZEAT étant trop faible. Le "coût" est le nombre total d'agglomérations ; il ne fait donc intervenir ni le nombre ni le coût unitaire des relevés qui font l'objet de la seconde optimisation. Notre principe consiste à dire que la variance intra-agglomération, laissée de côté pour l'instant, dépend du nombre total de relevés, et n'est pas affectée si on diminue seulement le nombre d'unités primaires en maintenant constant le nombre total de relevés. Par contre, le coût total de collecte ne peut qu'être diminué si on concentre les relevés dans moins d'agglomérations. Ainsi, on se permet de mettre sur un pied d'égalité les différentes CC. La valeur de référence $\tilde{V}$

est calculée à formule identique avec la répartition par CC-ZEAT actuelle. La dispersion $S^2(cc,z,v)$ par CC-ZEAT-Variété est estimée en remplaçant l'indice vrai d'agglomération-Variété par son estimateur. On intègre ainsi une part de la variance intra-agglomération pour chaque variété, ce qui implique que l'optimisation est en réalité effectuée sur "plus" que la variance inter-agglomérations. Cependant, comme la partie intra-agglomérations de la variance totale comporte toujours un facteur égal à l'inverse de la taille de l'échantillon d'agglomérations, ce procédé ne doit pas dégrader la variance totale. Lorsque la variété n'est relevée que dans une seule agglomération de la CC-ZEAT, on estime la dispersion $S^2(cc,z,v)$ par la moyenne des dispersions de la CC-variété sur l'ensemble des ZEAT (si on choisit d'utiliser la moyenne des dispersions de la CC-ZEAT sur l'ensemble des variétés, les résultats ne sont pas sensiblement modifiés). Si la pondération CC-ZEAT-variété est nulle, on n'effectue aucune correction (cela arrive quand le poids de la variété est faible). Cela équivaut à une repondération des variétés pour lesquelles l'information existe.

Comme lors du calcul de précision, on se heurte à des problèmes de définition des strates, puisque les agglomérations actuelles résultent du tirage de 1970. L'optimisation devant servir à l'échantillon futur, il nous a semblé plus juste d'utiliser l'information plus récente du recensement de 1982 : on considère ainsi que le nombre d'agglomérations tirables par CC-

ZEAT est le nombre d'agglomérations recensées en 1982 dans la CC-ZEAT, et on replace les agglomérations de l'échantillon dans la CC qui était en 1982 la leur ; l'estimateur des dispersions est biaisé, mais on peut espérer que le biais est faible, les changements de strate étant souvent dus à de faibles évolutions de population qui ne doivent pas modifier sensiblement la valeur numérique des indices de variété-agglomération.

Nous sommes contraints par le nombre total d'agglomérations tirables dans la CC-ZEAT. En réalité, on optimise dans un premier temps sans tenir compte de cette contrainte. Si le programme réclame, à l'optimum, davantage d'agglomérations qu'il n'en existe, alors on contraint la taille d'échantillon, dans la ou les CC concernées, à valoir l'effectif d'agglomérations disponibles, et on relance l'optimisation sur les autres strates.

Pratiquement, on a suivi la procédure suivante, pour chaque année, de 1981 à 1987 :

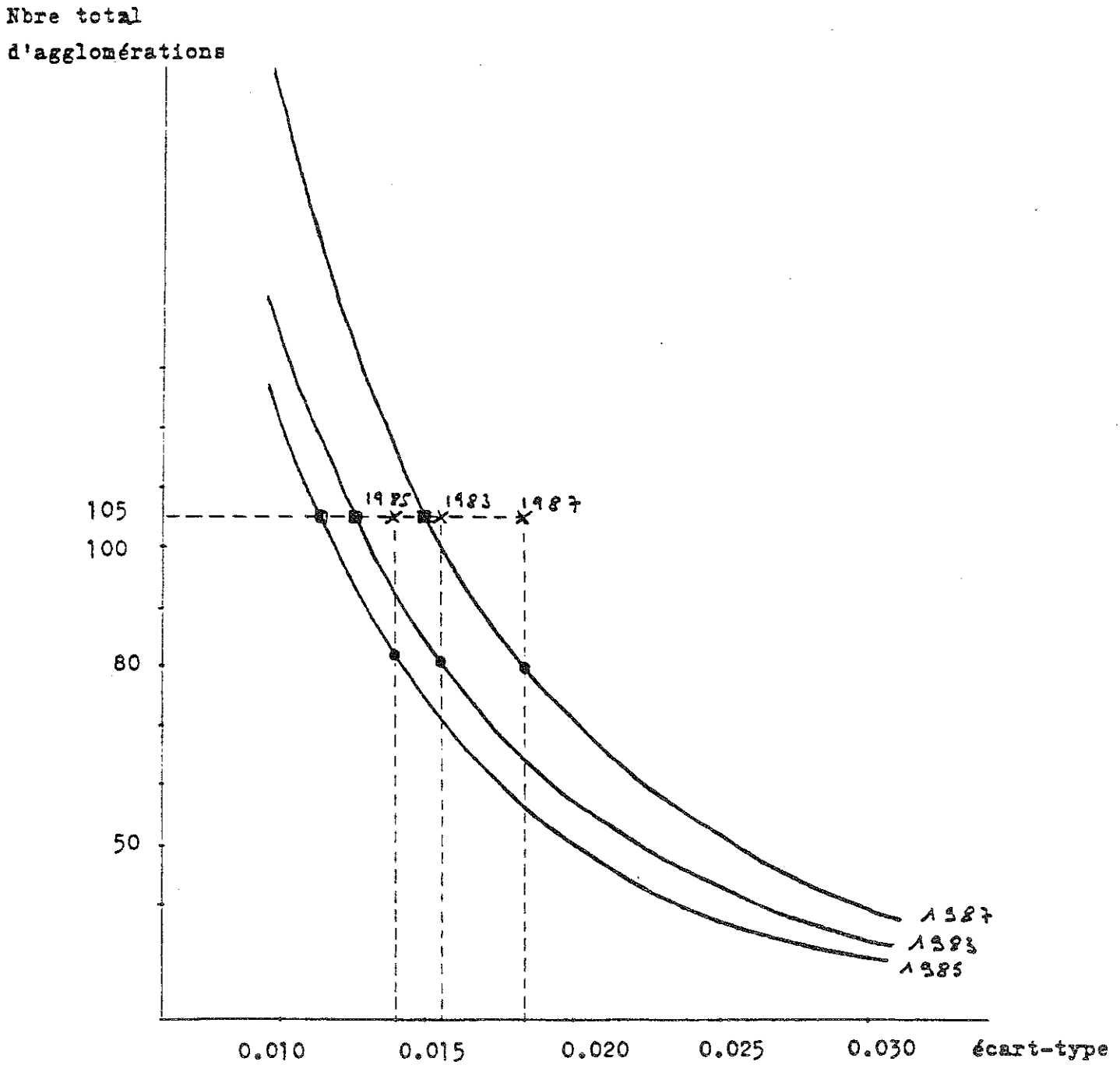
- On calcule la précision de l'indice avec l'expression de la variance utilisée dans le programme d'optimisation ( ). On obtient, pour 1987 par exemple, un écart-type de 0,018.

- On choisit neuf précisions a priori (entre 0,01 et 0,03) et on détermine pour chacune d'elles l'échantillon optimum par CC-ZEAT.

- A partir des résultats, on représente graphiquement la relation entre la précision et la taille totale de l'échantillon optimum (graphique 3).

Graphique 3

TAILLE DE L'ECHANTILLON OPTIMUM D'AGGLOMERATIONS
SUIVANT LA PRECISION VOULUE DE L'INDICE GENERAL



x échantillon et précision de l'indice actuels
● échantillon optimum sous contrainte de la précision de l'indice actuel
■ précision optimale sous contrainte de l'échantillon d'agglomérations actuel

P. Ardilly et F. Guglielmetti, Insee Méthodes n° 29-30-31

Les courbes ajustées obtenues pour chaque année (qui ont la forme hyperbolique attendue) sont remarquablement parallèles. Cela signifie que, quelle que soit la taille de l'échantillon, il y a de "bonnes" années (1985 par exemple) et de "moins bonnes" années (1987).

On lit sur le graphique la taille globale de l'échantillon correspondant à l'optimum sous la contrainte du maintien de la précision de l'indice estimée chaque année. Elle varie, selon l'année, entre 75 et 83 agglomérations (80 en 1987).

On répartit cet échantillon global par CC-ZEAT par interpolation à partir des résultats obtenus avec les précisions a priori. Le tableau 3 fournit la répartition par CC.

Compte tenu de l'imprécision des nombres obtenus par CC-ZEAT, il est plus prudent de proposer comme échantillon optimal une moyenne de plusieurs répartitions optimales annuelles. C'est ce qui est fait dans le tableau 4, qui rappelle en outre la structure de l'échantillon actuel.

Celui-ci sur-représente inutilement les agglomérations de moins de 200 000 habitants.Cependant l'optimisation ne bouleverse pas la répartition actuelle (il n'y a que la strate "Agglomérations de plus de 200 000 hab. du Midi Méditérranéen" où l'échantillon optimum est supérieur à l'échantillon actuel).

En pratique, l'échantillon optimum conserve, chaque fois que cela est possible, les agglomérations de l'échantillon actuel, pour des raisons évidentes d'organisation. Cette procédure n'est pas en contradiction avec la théorie statistique dans la mesure où ces agglomérations résultent bien d'un tirage aléatoire.

TABLEAU 3

Répartition optimale de l'échantillon d'agglomérations par catégorie de commune sous la contrainte du maintien de la précision de l'indice de l'année

Catégorie de commune	1981	1982	1983	1984	1985	1986	1987
2 000 - 10 000 hab.	14	12	14	10	14	14	11
10 000 - 100 000 hab.	30	26	30	29	29	29	33
100 000 - 200 000 hab.	13	14	13	14	15	16	13
+ 200 000 hab.	24	22	24	23	23	23	22
Paris	1	1	1	1	1	1	1
TOTAL	82	75	82	77	82	83	80

Tableau 4

OPTIMISATION DE L'ECHANTILLON D'AGGLOMERATIONS

ECHANTILLON ACTUEL (1989)

		ZEAT								
		1	2	3	4	5	7	8	9	Total
CC	2	1	6	-	4	5	3	3	3	25
	4	2	10	3	5	4	6	5	4	39
	6	-	2	4	2	4	1	2	2	17
	8/9	1	4	4	3	2	2	4	4	24
	Tot	4	22	11	14	15	12	14	13	105

ECHANTILLON OPTIMUM

		ZEAT								
		1	2	3	4	5	7	8	9	Total
CC	2	1	3	1	2	2	2	1	2	14
	4	2	7	2	3	3	5	4	3	29
	6	-	2	2	2	3	1	1	2	13
	8/9	1	3	4	3	2	2	4	5	24
	Tot	4	15	9	10	10	10	10	12	80

Echantillon optimum : moyenne des répartitions optimales obtenues pour les indices de 1981 à 1987.

ZEAT : 1 Région parisienne
2 Bassin parisien
3 Nord
4 Nord-Est
5 Ouest
7 Sud-Ouest
8 Centre-Est
9 Midi méditerranéen

CC : Catégorie de commune
2 Agglomérations de 2 000 à 10 000 hab.
4 Agglomérations de 10 000 à 100 000 hab.
6 Agglomérations de 100 000 à 200 000 hab.
8 Agglomérations de plus de 200 000 hab.
9 Agglomération de Paris

3.2. ECHANTILLON DE RELEVES

Une fois la liste des agglomérations déterminée, on désire répartir au mieux les relevés entre les agglomérations. Il est logique de raisonner conditionnellement à l'échantillon d'unités primaires, fixées pour plusieurs années. Les coûts de gestion liés aux réaffectations des relevés de variété ont été jugés admissibles, d'autant plus que notre variable de base est le nombre de relevés par variété-CC. Ces réaffectations seront réalisées chaque année en fonction des résultats obtenus les années antérieures. L'unité géographique élémentaire est certes l'agglomération, mais le nombre de relevés par variété-agglomération est trop faible (et le restera de toutes façons) pour que l'on puisse se baser sur des estimations de dispersion à ce niveau. En conséquence, en adoptant pour règle que, dans une CC, il y aura le même nombre de relevés dans chaque agglomération, on détermine la répartition des relevés au niveau variété-CC à partir de la moyenne des dispersions d'indices élémentaires, moyenne calculée sur l'ensemble des agglomérations de la CC. Cette règle recoupe la théorie en ce qui concerne les poids, identiques en principe dans chaque agglomération, et les coûts unitaires, fonction seulement de la CC et de la variété, mais ne prend pas en compte d'éventuelles différences de dispersion des indices élémentaires entre les agglomérations de la CC.

Ainsi, on veut :

$$MIN \sum_{cc,v} w^2(cc,v) \frac{s^2(cc,v)}{n(cc,v)}$$

$$sc \left| \sum_{cc,v} n(cc,v) \times \tilde{c}(cc,v) = \overline{C} \right.$$

où w(cc,v) est le poids du croisement cc x variété, n(cc,v) le nombre de relevés dans le croisement, $s^2(cc,v)$ l'estimateur de la dispersion calculé sur le modèle de ( ), $\tilde{c}(cc,v)$ le coût d'un relevé de (v) dans la cc, et $\overline{C}$ le budget total de l'enquête.

Les combinaisons variétés-CC qui ont un poids non nul mais une dispersion manquante sont prises en compte : on estime la dispersion par la moyenne des dispersions des CC renseignées pour la variété en question. Si le poids est nul, on ne peut rien dire, et le résultat final reste à valeur manquante : la détermination du nombre de relevés à effectuer restera entièrement liée à des contraintes de budget et de terrain (ce cas de figure survient surtout dans les petites agglomérations de la CC2, qui contribuent peu à la consommation nationale). Certaines variétés sont soumises à un traitement particulier : les tarifs nationaux ne nécessitent qu'un seul relevé, les tarifs locaux autant de relevés qu'il y a d'agglomérations, et les biens durables ne sont jamais relevés en CC 2. On s'oriente d'ailleurs actuellement vers une suppression des relevés d'habillement en CC 2, pour des raisons de difficulté d'observation.

La résolution du programme fournit des nombres réels qui sont arrondis selon un algorithme de répartition au plus fort reste. Pour se protéger contre les estimations aberrantes, on décide, si l'estimateur se base sur moins de X relevés, et si la valeur "optimale" résultante dépasse Y relevés, d'attribuer d'autorité un nombre moyen raisonnable de relevés dans la CC. Par exemple, on impose : 30 relevés en CC8 si X est inférieur à 10 et Y supérieur à 60.

L'inconvénient de ce type de programme est qu'il ne contrôle pas les précisions par poste : même si nous obtenons un gain de précision notable sur l'indice d'ensemble, certains postes se trouvent dégradés. On peut donc chercher à minimiser le coût global sous contrainte qu'une "bonne partie" des postes ait un écart-type inférieur à une certaine valeur. Dans un premier temps, on dresse la liste $\mathcal{S}$ des postes dont l'écart-type est supérieur à $\overline{\sigma}(p)$, paramètre d'entrée au choix. Puis on résoud :

$$MIN \sum_{p \in \mathcal{S}} \sum_{(cc,v) \in p} m(cc,v)\, \tilde{c}(cc,v)$$

$$sc \left| \forall p \in \mathcal{S} \sum_{(cc,v) \in p} w^2(cc,v|p) \times \frac{s^2(cc,v)}{m(cc,v)} \leq (\overline{\sigma}_{(p)})^2 \right.$$

où w(cc,v|p) est le poids du croisement cc x v dans le poste.

On détermine le coût optimum résultant de ce programme, puis, en le soustrayant au budget global, le coût restant à

affecter aux postes d'écart-type non contraint. Sur les postes non contraints, on réalise alors une optimisation du premier type (selon ( )). Malheureusement, si la contrainte est trop ambitieuse, on peut recueillir en fin de parcours un trop grand nombre de postes dont l'écart-type se trouve fortement détérioré : en effet, la première partie de l'optimisation a tendance à demander un nombre important de relevés supplémentaires, relevés qui font défaut aux postes qui n'étaient que moyennement bons à l'origine. Cette pratique a un intérêt parce que l'INSEE publie l'ensemble des indices de poste, indices dont on a vu la médiocre qualité. On préfère ainsi, dans cette optique, avoir un maximum d'indices "pas trop mauvais" (à considérer en fonction du budget), quitte à détériorer les bons indices, plutôt que de bons indices côtoyant de mauvais ou très mauvais indices. On admet implicitement (et cela peut être discuté) que deux indices quelconques sont dignes du même intérêt, alors que par ailleurs ils peuvent avoir des poids très différents.

Les calculs concernent la constitution de la nouvelle grille de répartition par CC-variété ; à partir de là, on a intérêt à répartir les relevés dans un nombre maximum d'agglomérations pour obtenir une variance minimale. Les contraintes de terrain interviennent alors pour imposer un nombre minimum de relevés par agglomération (soit deux actuellement). Si le nombre de relevés demandé dans le CC est trop faible, on ne pourra pas enquêter dans toutes les agglomérations de la CC ;

dans ce cas, on peut s'interroger sur la signification du premier échantillon de 105 agglomérations (qui va passer à 80) qui a fonction de "réserve optimale", et surtout sur le caractère aléatoire de la taille du sous-échantillon d'agglomérations dans lesquelles on effectue réellement des relevés. En effet, une fois déterminé le nombre d'agglomérations à enquêter, en appliquant une règle empirique, on s'orientera probablement vers une spécialisation des agglomérations par poste : ainsi, on cherchera à relever un maximum de variétés du poste dans la même agglomération, de façon à économiser des déplacements. Ce procédé ne devrait pas modifier sensiblement la précision, même si, à priori, on introduit un léger effet de grappe.

Dans le programme d'optimisation, les contraintes sur les coûts et la précision sont paramétrées, ce qui permet de faire des simulations avec une batterie de contraintes et de mesurer les effets de chacune d'elles.

Les programmes ont pour l'instant tourné sur les échantillons de l'indice de 1987 et 1988. Les résultats sont,au moins globalement,très voisins d'une année sur l'autre. Ceux qui sont présentés ici portent sur l'année 1987.

Une première série de simulations introduit comme contrainte de coût global le nombre total de relevés. On fait ici l'hypothèse implicite que le coût est le même pour chaque relevé. Cette hypothèse, manifestement erronée, permet cepen-

dant une mesure simple de l'influence de la taille de l'échantillon sur la précision et la structure de l'échantillon à l'issue de l'optimisation. Les principaux résultats obtenus sont les suivants (tableau 5) :

- l'échantillon actuel, s'il est optimisé, améliore très sensiblement la précision (l'écart-type global diminue de 40 %)
- on peut réduire de 2/3 l'échantillon sans détériorer la précision actuelle
- à coût constant, l'optimisation redistribue l'échantillon entre les secteurs selon le schéma suivant :

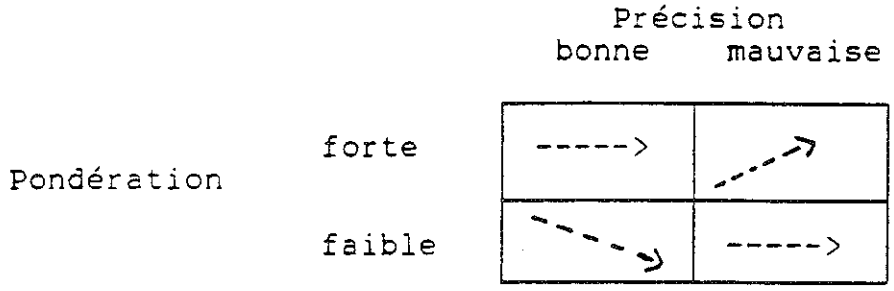


Tableau 5

OPTIMISATION DE L'ECHANTILLON DE RELEVES (1987)

Influence de la taille de l'échantillon

	Nombre total de relevés (milliers)					
	128 actuel (rappel)	128 optimum	100	75	50	30
ECART-TYPE PAR SECTEUR						
Alimentation	0.066	0.052	0.059	0.068	0.083	0.107
Habillement	0.149	0.095	0.108	0.124	0.151	0.195
Autres manufacturés	0.065	0.039	0.044	0.051	0.062	0.079
Services	0.100	0.043	0.049	0.055	0.066	0.084
Ensemble	0.044	0.025	0.028	0.033	0.040	0.051
STRUCTURE DE L'ECHANTILLON (milliers)						
Alimentation	43	35	27	20	13	8
Habillement	23	24	19	14	9	6
Autres manufacturés	43	40	31	24	16	9
Services	19	29	23	17	12	7
Ensemble	128	128	100	75	50	30

L'indice des Services n'est pas bon mais le poids correspondant est élevé ; l'optimisation augmente de 50 % le nombre de relevés, et la précision double. Les relevés sont pris au Secteur Alimentation qui pèse moins et dont la précision, bonne, souffre peu de la diminution de son échantillon. L'indice de l'habillement est mauvais ; une redistribution interne améliore la précision, mais compte tenu de sa petite pondération, l'effet induit par une augmentation de son échantillon serait faible sur la précision globale, ce qui explique que l'optimisation le laisse pratiquement inchangé.

Le coût d'un relevé mesure la plus ou moins grande difficulté à trouver, suivre ou remplacer l'article à observer ; il dépend de sa nature et du lieu d'observation (catégorie d'agglomération, type de point de vente) mais aussi des conditions de contrôle, de saisie et de traitement du prix. Il n'y a donc aucune raison qu'il soit identique pour tous les relevés ; mais sa mesure reste difficile.

L'optimisation a été faite en introduisant dans la contrainte de coût global plusieurs grilles alternatives de coûts différenciés par secteur et catégorie de commune.

Le tableau 6 illustre l'influence de ce type de contrainte sur la précision et la répartition de l'échantillon. Dans cet exemple, le coût d'un relevé est trois fois plus élevé dans l'habillement et les services que dans l'Alimentation, et

Tableau 6
OPTIMISATION DE L'ECHANTILLON DE RELEVES (1987)

Influence des coûts différenciés par secteur et catégorie de commune sous la contrainte du maintien du coût global actuel.
Echantillon (milliers)

	Catégorie de commune						Ecart-type
	2	4	6	8	9	T	
REPARTITION ACTUELLE							
Alimentation	5	10	5	10	13	43	0.066
Habillement	2	6	3	5	7	23	0.149
Autres manufacturés	3	11	7	11	11	43	0.065
Services	2	5	3	4	5	19	0.100
Ensemble	12	32	18	30	36	128	0.044
REPARTITION OPTIMUM (Coûts unitaires différenciés (*))							
Alimentation	4	10	6	12	18	50	0.045
Habillement	1	5	2	6	7	21	0.104
Autres manufacturés	2	9	5	11	17	44	0.038
Services	2	4	3	6	11	26	0.047
Ensemble	9	28	16	35	53	141	0.025

(*) Grille des coûts

Secteur \ CC	2	4	6	8	9
Alimentation	3	2	2	1	1
Habillement	9	5	5	3	3
Prod. manufacturés	5	3	3	2	2
Services	7	5	4	3	3

CC 2 agglomérations de 2 000 à 10 000 hab.
4 agglomérations de 10 000 à 100 000 hab.
6 " de 100 000 à 200 000 hab.
8 " de plus de 200 000 hab.
9 " de Paris

également trois fois plus élevé dans les agglomérations de moins de 10 000 habitants que dans celles de plus de 200 000 habitants. Le coût est mesuré en "unité de coût" et n'a qu'une valeur relative. Le nombre de relevés n'a plus qu'une importance secondaire et n'est plus contrôlé.

Les principaux résultats obtenus sont :

- Pour un même coût global, l'optimisation permet de faire 13 000 relevés supplémentaires

- Ces relevés bénéficient aux strates "bon marché" (Alimentation, grosses agglomérations), mais aussi au secteur des Services à cause de sa très mauvaise précision (il perd néanmoins des relevés par rapport à une optimisation avec un coût unitaire unique)

- Le choix de telle ou telle grille de coûts différenciés a peu d'influence sur les précisions par secteur. Dans le cas cité, la précision de l'indice général ne varie pas (0,025).

Contraindre les indices de poste à être bons (pour être publiables) va à l'encontre de la répartition optimale des relevés assurant la meilleure précision de l'indice général. Le tableau 7 montre très clairement que, pour un coût global donné

(mesuré ici simplement par le nombre total de relevés), une exigence croissante sur la précision des postes entraîne une diminution sensible de la précision des secteurs et de l'ensemble. Le phénomène est très net au seuil de 0,7 : il s'agit du mode de la distribution des écarts-types, et beaucoup de postes se trouvent brusquement parmi les postes à améliorer ; le transfert des relevés qui en résulte dégrade fortement les postes lourds qui assuraient la bonne qualité d'ensemble.

Tableau 7

OPTIMISATION DE L'ECHANTILLON DES RELEVES
AVEC UNE CONTRAINTE SUR LA PRECISION DES INDICES DE POSTE

	absence de contrainte (rappel)	Ecart-type intra (poste) inférieur à :			
		0,9	0,8	0,7	0,6
ECART-TYPE PAR SECTEUR					
Alimentation	0.052	0.055	0.057	0.060	0.087
Habillement	0.095	0.110	0.111	0.115	0.114
Autres manufacturés	0.039	0.047	0.046	0.051	0.062
Services	0.043	0.055	0.054	0.055	0.077
Ensemble	0.025	0.030	0.030	0.031	0.040
STRUCTURE DE L'ECHANTILLON (milliers)					
Alimentation	35	36	35	34	30
Habillement	24	24	26	29	37
Autres manufacturés	40	40	41	42	45
Services	29	28	26	23	16
Ensemble	128	128	128	128	128

4. CONCLUSION

Malgré la rareté de l'information auxiliaire (contrairement aux USA par exemple, dans un contexte méthodologique très différent), et une organisation des données peu adaptée au calcul de précision, on peut accorder une grande confiance aux précisions des indices de secteurs et à l'indice d'ensemble, ainsi qu'à l'effectif optimum de relevés par secteur-catégorie de commune.

Dans le cadre de la rénovation de l'indice des prix à la consommation, ces résultats ont finalement deux aspects : d'une part la connaissance nouvelle d'une information touchant ce point sensible qu'est la qualité de l'indice, d'autre part un aspect opérationnel. En effet, l'optimisation des échantillons, couplée avec les estimations de précision, permet, outre la mise à jour annuelle des effectifs, la remise en cause de la nomenclature des postes et du choix des variétés sur des bases statistiques.

Cependant, on ne peut prendre en compte les précisions des postes pour l'optimisation que si l'on dispose de plusieurs années.

La seule opération qui a pu être menée sur des données antérieures à 1987 est une estimation par bootstrap sur les indices estimés d'agglomérations pour l'ensemble des produits entre 1981 et 1987. Cette estimation est trop grossière pour fournir des résultats comparables à ceux de la méthode

analytique, mais suffisante pour montrer que la précision de l'indice d'ensemble ne dépend pas du niveau de l'inflation. Ce résultat, intéressant en soi, permet d'espérer une certaine stabilité dans le temps des résultats de l'optimisation.

5. REFERENCES

- Pour comprendre l'indice des prix.INSEE avril 87 (2e édition)
- Bulletin mensuel de statistique. INSEE
- Item-outlet Sample redesign for the 1987 US consumer price index revision. US Bureau of labor statistics.
- Cochran W.G. Sampling techniques. Wiley N.Y. 1977
- Divers travaux présentés au Séminaire sur les statistiques des prix à la consommation. Genève. juin 1986