Calcul de l'échantillon d'agglomérations optimal pour l'indice des prix suite au passage aux données de caisses

Sébastien FAIVRE[1]

[1] *INSEE, sebastien.faivre@insee.fr*

I. Le tirage des agglomérations de l'échantillon actuel

Le champ actuel de l'indice des prix porte sur l'ensemble des agglomérations de plus de 2000 habitants. Chaque mois, environ 160 000 prix sont relevés par les enquêteurs dans plus de 27 000 points de ventes.

La collecte est réalisée dans un échantillon de **96 agglomérations**, stratifié par groupes de région (ZEAT) et par tranches de taille d'unités urbaines. Il existe **8 ZEAT** (Région parisienne, bassin parisien, Nord, Est, Ouest, Sud-Ouest, Centre-est, Méditerranéen) et **4 tranches de tailles d'unités** urbaines (2000 à 20 000 habitants, 20 000 à 100 000 habitants, plus de 100 000 habitants, Paris), **notées généralement CC** (catégories de communes) dans les articles méthodologiques.

L'échantillon actuel de 96 agglomérations est le fruit de l'histoire et des différentes générations d'indices de prix, qui ont été l'occasion de « mises à jour » successives de l'échantillon initial d'agglomérations tiré pour l'indice en base 1962.

Le premier tirage d'agglomérations (87 agglomérations) a eu lieu pour l'indice en base 1962 publié à partir de janvier 1962 (auparavant, la collecte était organisée uniquement sur Paris dans les 17 villes siège des Directions Régionales de l'INSEE). L'échantillon d'agglomérations a été élargi à 108 lors du passage à l'indice en base 1970 publié à partir de 1971. Par la suite, l'échantillon d'agglomérations a été mis à jour en 1992 dans le cadre du passage à l'indice en base 1990, à l'issue d'un travail d'optimisation de l'échantillon mené par Pascal Ardilly et Francis Guglielmetti (cf. infra paragraphe III, l'optimisation de l'échantillon pour l'indice en base 1990).

Compte-tenu de l'ancienneté du tirage, on ne dispose pas d'informations sur le mode de sélection des 87 agglomérations pour la base 1962.

En revanche, le principe du tirage mis en œuvre pour la sélection des 108 agglomérations retenues pour l'indice en base 1970 est décrit dans l'article de Pascal Ardilly et Francis Guglielmetti « Optimisation de l'échantillon pour le calcul de l'indice de prix à la consommation » publié dans les actes des JMS de 1991 (INSEE Méthodes 29-30-31, pages 74 et 75).

Le tirage des unités primaires s'effectue de la manière suivante : considérant l'ensemble des agglomérations de plus de 2000 habitants, on réalise une stratification préalable par CC[1]-ZEAT. On retient d'office toutes les agglomérations de CC8 [plus de 200 000 habitants], ainsi

[1] CC=Tranche de taille d'agglomérations

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.

Ce mode de sélection **constitue un compromis entre un tirage théorique des agglomérations proportionnel à leur taille en terme de nombre d'habitants** (en divisant chaque strate en sous-groupes définis en fonction de la taille de l'agglomération) **et la nécessité pour des raisons pratique de limiter le renouvellement de l'échantillon d'agglomérations** (en conservant autant que possible dans chaque groupe les agglomérations déjà présentes).

Choix d'un plan de sondage des agglomérations proportionnel à leur taille

L'indice global des prix qu'on cherche à estimer est la moyenne pondérée (par leur poids en nombre d'habitants) des indices des prix observés dans chaque agglomération.

$$I = \sum_{agglo} w_i I_i$$

Cet indice est estimé sur l'échantillon d'agglomérations par :

$$\hat{I} = \sum_{k \in S} \frac{w_k}{\pi_k} I_k$$

Si la probabilité de sélection de l'agglomération k est proportionnelle à sa taille en terme de nombre d'habitants, alors $\hat{I}$ s'écrit comme la moyenne arithmétique simple des indices d'agglomération :

$$\hat{I} = \frac{1}{n} \sum_{k \in S} I_k$$

Cela permet ainsi de limiter la dispersion des poids et d'obtenir des estimateurs plus robustes.

Ce mode de sélection des agglomérations a été reconduit pour le passage en base 1990.

Préalablement, un travail d'optimisation de l'échantillon d'agglomération avait été effectué par Ardilly et Guglielmetti, aboutissant au calcul du nombre minimal d'agglomérations à tirer dans chaque strate *TrancheUnitésUrbainesxZEAT* pour obtenir un niveau de précision donné pour le tirage de premier degré (phase détaillée dans la deuxième partie de la note).

Dans le cadre méthodologique exposé ci-dessus, la procédure pratique de sélection des 96 agglomérations de l'échantillon en base 1990 est décrite dans une note de synthèse sur la construction de l'indice des prix lors de la rénovation pour la base 1990 rédigée par Sandra Montiel en 1995 (note 391/F320 du 27 novembre 1995)

Dans un premier temps, on découpe la France en strates. Chaque strate est composée de l'ensemble des agglomérations d'une catégorie de communes donnée, appartenant à une ZEAT donnée. (…). Les strates de tailles sont divisées en sous-groupes plus homogènes afin d'affiner la représentativité de l'échantillon :

- 2 000 à 20 000 habitants, 4 sous groupes
- 20 000 à 100 000 habitants, 6 sous groupes
- + de 100 000 habitants, 2 sous groupes
- Paris, 1 sous groupe

Théoriquement, le nombre d'agglomérations à retenir dans chaque catégorie de communes est proportionnel aux taux de représentativité choisis (variables d'une catégorie de commune à une autre), et identique pour toutes les ZEAT. Des taux plus élevés sont retenus pour les petites agglomérations afin de se rapprocher facilement d'un échantillon auto pondéré d'observations.

(…)

Le tirage agglomération accorde le même poids dans une taille, hormis le cas où l'agglomération a une probabilité d'inclusion de 1 (cas des grandes métropoles et de Paris), elle a alors son poids réel dans la taille. Le poids est corrigé à la marge a posteriori, afin d'avoir dans l'indice le poids réel de chaque région.

Ce tirage est tel qu'il permet de conserver, sans trop de biais, un nombre important d'agglomérations de l'ancien échantillon. On tire des tranches fines de catégories d'agglomérations et on représente ces tranches par le plus grand nombre d'agglomérations déjà utilisées. Cependant, il faut faire attention à l'évolution des villes, elles peuvent passer d'une strate à l'autre entre deux recensements. De plus, la structure aussi peut varier : une plus grande part de la population peut vivre et consommer dans de petites agglomérations. Il faut que la représentativité soit la meilleure possible au moment du tirage.

(…)

Un groupe pour une région donnée correspond à une unité primaire du tirage à 2 degrés. Ces unités primaires sont tirées avec une probabilité inégale. Le tirage est systématique pour les groupes à l'intérieur d'une strate donnée (ZEAT et catégorie) avec des probabilités proportionnelles à l'effectif de chaque groupe.

Les unités secondaires sont les agglomérations de chaque groupe. Un groupe est tiré n fois (n peut être nul), on tire alors n agglomérations dans ce groupe.
Soit m le nombre d'agglomérations dans le groupe de l'échantillon optimisé.

- Si m est inférieur à n, on garde les m agglomérations et on tire avec des probabilités égales les n-m agglomérations nécessaires

- Si m est supérieur à n, on tire n agglomérations parmi les m

II. L'opération d'optimisation de l'échantillon actuel d'agglomérations

A. Le cadre théorique retenu pour le calcul de la précision

Le cadre théorique retenu est celui d'un tirage à deux degrés avec un sondage aléatoire simple à chaque degré de tirage (tirage des agglomérations au sein de chaque strate, puis tirage des points de ventes dans lesquels les relevés sont effectués au sein de chaque agglomération).

Dans le cadre des simplifications nécessaires à la modélisation du plan de sondage, Ardilly et Guglielmetti précisent que le système de tirage de groupes n'est pas pris en compte, car, « pour certaines variétés, l'échantillon d'agglomérations enquêtées est trop faible : plutôt que de manipuler un échantillon stratifié avec une 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 (2 000 à 10 000 habitants), 4 (10 000 à 100 000 habitants) et 6 (100 000 à 200 000 habitants), la taille d'échantillon était fixée et égale à celle qui avait été déterminée lors de l'affectation des poids. Pour la CC8 (plus de 200 000 habitants), 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. ».

Les indices étudiés sont ceux de l'inflation annuelle, c'est-à-dire pour une année donnée l'indice issu de la comparaison entre le mois 0 (mois de décembre de l'année précédente) et le mois 12 (mois de décembre de l'année considérée).

B. Le programme d'optimisation du nombre d'agglomérations par strate

La variance de l'indice des prix est calculée comme la somme de la variance de première phase (liée au tirage des agglomérations) et de la variance de seconde phase (liée au tirage des points de ventes au sein de chaque agglomération de l'échantillon).

Pour les strates non exhaustives (agglomérations de moins de 200 000 habitants), l'indice des prix d'une variété v donnée au sein d'une tranche d'agglomération cc et d'une ZEAT z 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 v est observée.

La variance totale de l'indice variété-strate $V(\hat{I}(cc,z,v))$ est la somme de la variance de première phase (associée au tirage des agglomérations au sein de la strate) $V^{1P}(\hat{I}(cc,z,v))$ et de la variance de seconde phase (associée au tirage des points de ventes au sein des agglomérations) $V^{2P}(\hat{I}(cc,z,v))$

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

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 v 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 v, $M(cc,z)$ 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.

L'estimateur « naïf » de la variance de première phase pour les indices variétés-strates est :

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

où $s^2(cc,z,v)$ est l'estimateur empirique de la dispersion des indices var-agglo sur l'échantillon d'agglomérations tirées dans la strate.

Bien que cet estimateur soit biaisé et surestime la vraie valeur de la variance inter-agglomérations, dans la mesure où il intègre également une partie de la variance intra-agglomérations (ce point est détaillé en annexe 2), il est utilisé par la suite dans les programmes d'optimisation pour des raisons de simplicité.

L'indice total est calculé comme : $\hat{I} = \sum_{cc,z,v} w(cc,z,v)\hat{I}(cc,z,v)$

Sous l'hypothèse d'indépendance des tirages entre les variétés et les strates, on a alors :

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

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

On met ensuite en œuvre un programme de minimisation pour calculer le nombre minimal d'agglomérations à tirer dans chacune des strates pour obtenir la même variance de premier degré que dans l'échantillon d'agglomérations de la base 1970[2].

[2] A noter que le niveau de variance de première phase associé à l'échantillon de la base 1970 est lui estimé par Ardilly et Guglielmetti avec l'estimateur sans biais de la variance de premier degré.

Le coût associé à un échantillon d'agglomérations est le nombre total d'agglomérations de l'échantillon : $C = \sum_{cc,z} m(cc,z)$. Il est à noter que ce coût est indépendant de la taille des agglomérations retenues[3].

Par ailleurs, on fait au niveau de l'optimisation l'hypothèse très simplificatrice que toutes les variétés sont relevées dans toutes les agglomérations d'une strate dès lors qu'elles sont relevées dans au moins une des agglomérations de cette strate[4].

On résout donc le programme de minimisation suivant :

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

sous contrainte

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

où $\tilde{V}$ est la variance de premier degré calculée avec l'échantillon d'agglomérations issu du tirage de 1970, sur la base de l'estimateur sans biais de la variance de premier degré.

C. Les résultats de l'optimisation pour le changement de base 1990

L'échantillon d'agglomérations en vigueur en 1989 (issu du tirage de 1970) comportait 103 agglomérations. Les calculs effectués sur la base du Recensement de 1982 montraient que, pour un même niveau de précision, l'échantillon optimal comportait 81 agglomérations.

Par rapport à l'échantillon optimal, l'échantillon en vigueur surreprésentait toutes les tranches d' agglomérations « non exhaustives » , et tout particulièrement les petites agglomérations (2 000 à 10 000 habitants) : 24 de ces agglomérations étaient suivies dans l'échantillon en vigueur contre 13 dans l'échantillon optimal.

Au final, l'optimisation sur la base du Recensement de 1990 a réduit ces écarts, puisque le nombre d'agglomérations finalement conservées était de 96 sur 101[5], compte tenu notamment du fait que la population des petites agglomérations (qui étaient surreprésentées dans le tirage 1970) avait augmenté entre le recensement de 1982 et celui de 1990.

Cependant, malgré l'écart plus faible entre l'échantillon d'agglomérations en vigueur et l'échantillon optimal calculé à partir du RP 1990, le renouvellement de l'échantillon d'agglomérations est resté relativement important puisque à l'issue de la procédure de tirage

[3] Ce qui tendrait plutôt à faciliter la sélection de « grandes » agglomérations avec un poids démographique plus important, puisqu'elles représentent une part plus importante de la variance totale que les petites agglomérations sans pour autant coûter plus cher à enquêter.

[4] Ce qui est loin d'être le cas en pratique, puisqu'une variété est relevée en moyenne dans 21 agglomérations sur 96.

[5] Au lieu de 103, 2 agglomérations de l'IPC ayant été absorbées une autre agglomération IPC à l'issue du recensement de 1990.

par groupes, 24 agglomérations de l'échantillon de 1970 ont été supprimées et 19 nouvelles agglomérations ont été créées[6].

III. Le calcul de l'échantillon optimal d'agglomérations après le passage aux données de caisses

A. Quelques ordres de grandeur sur l'impact potentiel du passage aux données de caisse sur la charge de travail du réseau enquêteurs

Sur la période octobre 2008 - septembre 2009, on a observé les coûts de collecte suivants :

	Alimentaire	Biens durables	Habillement	Manufacturé	Services	Total
Nombre total de relevés	413 105	77 857	246 908	366 999	238 845	1 343 714
Dont GMS	300 123	21 929	41 632	139 798	3 075	506 557
Rémunération unitaire relevé (hors charges patronales)	0,82	1,28	1,76	1,28	1,28	
Rémunération totale relevés, hors charges patronales (euros)	338 746	99 657	434 558	469 759	305 722	1 648 441
Dont rémunération relevés GMS (euros)	246 101	28 069	73 272	178 941	3 936	530 320

On constate ainsi qu'un tiers (33%) des relevés de prix effectués par les enquêteurs le sont dans les grandes et moyennes surfaces et concernent l'alimentaire et les produits manufacturés, secteurs qui pourraient être suivis au moyen des données de caisses.

Par ailleurs, la rémunération « directe » des relevés de prix « hors produits frais » est estimé à 1,65 millions d'euros par an, dont 0,42 millions d'euros pour les relevés effectués dans les grandes et moyennes surfaces dans les secteurs de l'alimentaire et des biens durables. Ces relevés représentent ainsi un peu plus d'un quart (26%) de la rémunération des enquêteurs.

Compte-tenu de la baisse importante des volumes de collecte qui serait associée au passage aux données de caisses, l'organisation de la collecte restante doit être repensée. Dans ce cadre,

[6] Le nombre d'agglomérations supprimées ou créées est calculé en comparant la liste des agglomérations présentes dans l'échantillon 1993 avec la liste des agglomérations présentes dans l'échantillon 1991, l'année 1992 (année de double collecte) étant une année de transition entre les deux échantillons.

une concentration de la collecte restante dans un nombre restreint d'agglomérations est étudiée ici.

B. Calcul du nombre d'agglomérations à conserver après le passage aux données de caisses

1. Principe du calcul du nombre optimal d'agglomérations à conserver après le passage aux données de caisses

L'objectif étudié ici est de conserver après le passage aux données de caisses le même niveau de précision pour l'indice global qu'avant le passage aux données de caisses[7].

On se contentera d'étudier ici la précision de l'indice sur le champ couvert par la collecte enquêteurs (pour laquelle on dispose de la modélisation de Ardilly/Guglielmetti), puisque le champ « hors collecte enquêteurs » n'est pas impacté par le passage aux données de caisses, et compte tenu également de 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…) qui rendent difficile d'effectuer un calcul de précision sur le domaine « hors collecte enquêteurs ».

2. Le cadre théorique du calcul

On reprend ici le cadre théorique « simplifié » proposé par Ardilly et Guglielmetti pour la calcul de la précision de l'indice courant, c'est-à-dire qu'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[8].

Enfin, on fait l'hypothèse que les échantillons de prix observés pour des deux variétés distinctes sont « indépendants ». Cette hypothèse est bien sûr discutable compte-tenu du fait que l'enquêteur effectue pour des raisons d'efficacité le plus grand nombre de relevés possibles dans un même point de vente. On reprend ici le choix effectué par Ardilly et Guglielmetti de négliger l'effet de grappe lié au tirage des points de ventes[9].

Lorsqu'une variété est relevée dans une seule agglomération dans une tranche d'agglomération (ce qui ne permet pas d'estimer la dispersion des évolutions de prix de la variété au niveau de la tranche d'agglomération), on impute alors la valeur de la dispersion

[7] On suppose que, compte-tenu de la possibilité de concentrer la collecte dans un nombre restreint d'agglomérations C et D sans diminuer le nombre d'agglomérations de relevés pour l'essentiel des variétés, la précision au niveau poste ne sera pas diminuée après le passage aux données de caisses (à l'exception de l'essence, des repas au restaurant et de la réparation mécanique, qui doivent faire l'objet d'une réflexion spécifique).

[8] On suppose ainsi que le nombre de points de ventes est « très grand » au regard du nombre de relevés effectués, ce qui peut être discuté, notamment dans les plus petites agglomérations où il a parfois été nécessaire d'ajouter une agglomération binôme pour effectuer tous les relevés nécessaires. Une telle approximation conduit ainsi à surestimer la variance de seconde degré, et donc la variance totale.

[9] Pour ce qui est de la covariance de premier degré, les calculs effectués par Ardilly et Guglielmetti montrent qu'elle est faible au regard de la variance, et qu'elle peut donc être négligée.

des évolutions de prix toutes tranches d'agglomérations confondues[10]. Il s'agit cependant d'un nombre de cas très réduit et concernant par définition des variétés dont le poids est faible dans la consommation, au mois dans la tranche d'agglomération où une seule agglomération est enquêtée.

En revanche, il n'existe pas de solution très satisfaisante au cas des variétés qui ne sont pas enquêtées dans certaines tranches d'agglomérations. Ce cas de figure, peu fréquent dans l'échantillon en base 1970 qui avait servi de base aux travaux d'Ardilly et Guglielmetti, était traité au niveau d'agrégation supérieur (niveau poste) soit en repondérant les variétés enquêtées du poste, soit en imputant aux variétés non enquêtées la précision moyenne des variétés enquêtées du poste.

Une telle procédure est bien adaptée à des imputations en nombre limité. Or, depuis le changement de base 1990, l'habillement n'est, pour des raisons de coût, (quasiment)plus collecté dans les agglomérations C et D (moins de 100 000 habitants), alors que c'était le cas auparavant. De même, les biens durables ne sont plus collectés dans les agglomérations D (moins de 20 000 habitants).

Ce « trou de collecte » créé ainsi un biais dans l'estimation de l'indice des prix, qu'il est difficile compte tenu de son ampleur de traiter par imputation (le risque étant sinon d'introduire des erreurs de mesure importantes). On choisit donc plutôt d'optimiser notre échantillon d'agglomérations sur le champ de la collecte effective, qui ne comprend pas l'habillement dans les agglomérations C (20 000 à 100 000 habitants) et l'habillement et les biens durables dans les agglomérations D (2 000 à 20 000 habitants).

On notera cependant qu'une telle procédure, qui est bien adaptée à l'optimisation de la collecte telle qu'elle est effectivement réalisée, tend à diminuer dans les calculs le poids et la variabilité associée aux agglomérations C et D, et donc à diminuer le nombre optimal d'agglomération à sélectionner dans ces tranches d'agglomérations par rapport à un calcul où le poids économique des secteurs « non couverts » dans ces agglomérations aurait été pris en compte.

3. Modélisation du gain de variance associé au champ couvert par les données de caisses

Le gain associé au passage aux données de caisses intervient aux deux degrés du plan de sondage.

- ***Au premier degré (tirage des agglomérations***), on peut considérer, compte-tenu des remontées de données effectuées par Nielsen et IRI, que l'ensemble du territoire sera couvert par les données de caisses. En particulier, il ne sera plus nécessaire de limiter la collecte à un échantillon d'agglomération comme c'est le cas avec la collecte enquêteur actuelle.

On fait donc ici l'hypothèse que les données de caisses permettront de couvrir soit la totalité des agglomérations du territoire, soit une si grande proportion, que la variance de premier degré sera (quasiment) réduite à zéro sur le champ des données de caisses.

[10] Ardilly et Guglielmetti utilisent une procédure d'imputation plus élaborée, cf. article des JMS 1991. Compte-tenu du nombre de cas très limité, une procédure d'imputation simple a semblé suffisante.

On prend donc dans toute la suite une hypothèse de variance de premier degré sur le champ des données de caisses égale à zéro.

- ***Au second degré de tirage (tirage des prix observés au sein des agglomérations)***, la variance est fonction du nombre de séries suivies.

De ce fait, l'évolution de la variance de second degré sur le champ des données de caisses dépendra de l'évolution du nombre de séries de prix suivies par rapport au panier actuel.

Dans un premier scénario (S1), on suppose qu'on conserve le même nombre de séries de produits suivis par agglomération que dans le panier actuel. Dans ce cas, la variance de second degré sur le champ des données de caisses reste la même qu'actuellement.

Dans un second scénario (S2), on suppose qu'on multiplie par deux le nombre de séries suivies par agglomération. Dans ce cas, la variance de second degré sur le champ données de caisses est divisée par deux par rapport à sa valeur actuelle.

Dans un troisième scénario (S3), on suppose qu'on arrive à suivre dans le panier représentatif l'ensemble des produits vendus en grandes et moyennes surfaces. Dans ce cas, la variance de second degré sur le champ données de caisses est réduite à zéro

4. Prise en compte du gain de variance sur le champ données de caisses pour l'optimisation de la collecte restante

On a avant le passage aux données de caisses un niveau de variance de l'indice des prix associé à la collecte $V_{actuel}(\hat{I})$, somme de la variance de première phase (tirage des agglomérations) $V^{1P}_{actuel}(\hat{I})$ et de la variance de seconde phase (tirage des prix au sein des agglomérations) $V^{2P}_{actuel}(\hat{I})$:

$$V_{actuel}(\hat{I}) = V^{1P}_{actuel}(\hat{I}) + V^{2P}_{actuel}(\hat{I})$$

Après le passage aux données de caisses, on a la variance suivante associée à la collecte (sous l'hypothèse d'une variance de première phase nulle sur le champ des données de caisses) :

$$V_{après}(\hat{I}) = V^{1P}_{après}(\hat{I}) + V^{2P}_{après}(\hat{I}) = V^{1P}_{après,HorsCaisses}(\hat{I}) + V^{2P}_{après,HorsCaisses}(\hat{I}) + \tilde{V}^{2P}_{après,Caisses}(\hat{I})$$

où la valeur de la variance de seconde phase sur le champ des données de caisses $\tilde{V}^{2P}_{après,Caisses}(\hat{I})$ varie dans les trois scénarios S1, S2 et S3 décrits au paragraphe précédent.

Comme on souhaite garder le même niveau de précision globale sur le champ actuellement couvert par la collecte avant et après le passage aux données de caisses, on a :

$$V_{après}(\hat{I}) = V_{actuel}(\hat{I})$$

Soit $V^{1P}_{après,HorsCaisses}(\hat{I}) + V^{2P}_{après,HorsCaisses}(\hat{I}) + \tilde{V}^{2P}_{après,Caisses} = V^{1P}_{actuel}(\hat{I}) + V^{2P}_{actuel}(\hat{I})$

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 : $V^{2P}_{après,HorsCaisses} = V^{2P}_{actuel,HorsCaisses}$. (remarque : 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).

On obtient au final :

$$V^{1P}_{après,HorsCaisses}(\hat{I}) = V^{1P}_{actuel}(\hat{I}) + V^{2P}_{actuel}(\hat{I}) - V^{2P}_{actuel,HorsCaisses} - \tilde{V}^{2P}_{après,Caisses}$$

Remarque :

On rappelle que $\tilde{V}^{2P}_{après,Caisses}(\hat{I}) \leq V^{2P}_{avant,Caisses}(\hat{I})$ car le nombre de séries par agglomération suives sur le champ des données de caisses après le passage aux données de caisses sera supérieur ou égal au nombre de séries par agglomération suivies actuellement.

En effectuant l'approximation $V^{2P}_{actuel}(\hat{I}) = V^{2P}_{actuel,Caisses}(\hat{I}) + V^{2P}_{actuel,HorsCaisses}(\hat{I})$ (ce qui n'est pas tout à fait exact en pratique car les indices de variétés sont calculés avec des formules non linéaires), on obtient :

$$V^{2P}_{actuel}(\hat{I}) \geq \tilde{V}^{2P}_{après,Caisses}(\hat{I}) + V^{2P}_{actuel,HorsCaisses}(\hat{I})$$

soit donc $V^{1P}_{après,HorsCaisses}(\hat{I}) \geq V^{1P}_{actuel}(\hat{I})$

On voit ainsi que, dans le cadre du passage aux données de caisses, on optimise l'échantillon d'agglomérations pour la collecte restante sous une contrainte moins stricte que pour la collecte actuelle (puisque la variance de première phase à atteindre est plus élevée que celle associée à la collecte actuelle) et sur un champ de collecte plus réduit, ce qui permettra de sélectionner un nombre d'agglomérations plus faible toutes choses égales par ailleurs.

5. Paramètres méthodologiques du calcul d'optimisation

Dans le contexte spécifique du passage aux données de caisses, la méthode mise en œuvre ici diffère de celle proposée par Ardilly et Guglielmetti sur deux points :

A. Niveau d'optimisation du nombre d'agglomérations

Pour le changement de base 1990, Ardilly et Guglielmetti avaient calculé un nombre d'agglomérations optimal au niveau *TrancheAgglomération*X*ZEAT*.

Compte-tenu de l'importance des hypothèses effectuées et du faible nombre d'agglomérations retenues en strates C et D après le passage aux données de caisses, on a préféré se limiter ici à calculer un nombre optimal d'agglomérations au niveau *TrancheAgglomération,* sans faire intervenir la dimension ZEAT.

B. Choix des estimateurs de variance interagglomérations et intraagglomérations

Il existe deux estimateurs pour la variance de premier degré : un estimateur « naïf » (obtenu directement à partir de l'expression de la vraie variance en remplaçant la vraie valeur de la dispersion interagglomération par son estimation à partir de l'échantillon) et un estimateur sans biais obtenu en retranchant à l'estimateur « naïf » un terme fonction de la dispersion intraagglomération.

Ces deux estimateurs sont détaillés en annexe.

L'estimateur de variance utilisé dans le programme d'optimisation pour le calcul du nombre optimal d'agglomérations à retenir est l'estimateur « naïf », l'estimateur sans biais faisant intervenir un terme de variance « intra-agglomérations » difficile à prendre en compte dans le cadre de l'optimisation. De ce fait, pour des raisons d'homogénéité et d'interprétation des résultats, la cible de variance a été calculée en prenant les estimateurs biaisés de la variance interagglomérations et de la variance intraagglomérations[11] (on rappelle que l'estimateur de la variance totale obtenu en sommant les estimateurs naïfs des variances interagglomérations et intraagglomérations est lui sans biais, cf. annexe 2).

6. Formule de calcul du nombre d'agglomérations optimal

Pour déterminer le nombre d'agglomérations à retenir par tranches d'agglomérations, on résout donc le programme d'optimisation suivant sur l'ensemble du champ de la collecte hors données de caisses :

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

sous contrainte

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

En posant $g^2(cc) = \sum_z \sum_v w^2(cc,z,v)s^2(cc,z,v)$, on obtient après résolution par la méthode du Lagrangien (cf. résolution en annexe) :

[11] A la différence d'Ardilly et Guglielmetti, qui utilisent l'estimateur sans biais de la variance de première phase pour calculer la cible de variance à atteindre (variance de première phase de l'échantillon en vigueur au moment du calcul d'optimisation).

$$m(cc) = \frac{g(cc)\sum_{CC} g(cc)}{V^{1P}_{après}(\hat{I}) + \sum_{CC} \frac{g^2(cc)}{M(cc)}}$$

7. Calculs d'optimisation effectués

Les calculs d'optimisation ont été effectués sur les données d'inflation annuelle 2009 (variations de prix entre décembre 2008 et décembre 2009), à partir d'une table spécifique produite par la chaîne IPCNAT pour le calcul de la précision de l'indice.

Les données nécessaires au calcul correspondant aux années antérieures (2003 à 2008) n'ont pas été archivées par la chaine IPCNAT (seules des tables donnant des résultats plus agrégés sont conservées), mais pourront être reconstituées par l'équipe informatique IPC fin 2010 dans le cadre des opérations de migration hors MVS, ce qui permettra alors d'effectuer également le calcul sur les années antérieures à 2009.

C. Résultats observés pour l'échantillon optimal d'agglomérations sur la base de la stratification d'agglomérations actuelle

Les calculs d'optimisation ont été effectués sur la base du découpage actuel en quatre tranches d'agglomérations :

A : Paris
B : agglomérations de plus de 100 000 habitants
C : agglomérations entre 20 000 et 100 000 habitants
D : agglomérations de mois de 20 000 habitants

L'agglomération de Paris est bien entendu sélectionné automatiquement dans l'échantillon d'agglomération optimal, ce qui fait que l'optimisation porte uniquement sur les tranches B, C et D.

Les résultats obtenus sont présentés dans le tableau ci-dessous :

Tableau 1 : nombre d'agglomérations optimal (hors agglomération Parisienne) avant et après le passage aux données de caisses dans le cadre du découpage actuel en 4 tranches d'agglomérations.

Tranche d'agglomération	Collecte actuelle		Collecte après passage aux données de caisses		
	Nombre actuel	Nombre optimal collecte actuelle	S1 : Variance de première phase égale 0 sur le champ données de caisses	S2 : Variance de première phase égale 0 et variance de seconde phase divisée par deux sur le champ données de caisses	S3 :Variance égale à zéro sur le champ données de caisses
B (+ 100 000 habitants)	37	46	43	36	31
C (20 000 à 100 000 habitants)	25	17	14	12	10
D (2 000 à 20 000 habitants	33	24	22	18	16

On constate ainsi que l'échantillon d'agglomérations IPC actuel sous-représente les agglomérations de plus de 100 000 habitants (agglomérations B) et surreprésente les agglomérations de moins de 100 000 habitants (agglomérations C et D). Ce phénomène s'explique en partie par le fait que l'échantillon d'agglomérations actuel avait été optimisé avec une collecte comprenant des relevés d'habillement dans les agglomérations C et D qui ont été supprimés par la suite.

Concernant le passage aux données de caisses, le tableau 2 ci-dessous montre que l'échantillon d'agglomérations envisagé pour des raisons de viabilité de la collecte (maintien des 37 agglomérations B, passage de 25 à 17 agglomérations C et de 33 à 20 agglomérations D) est compatible avec la maintien de la précision actuelle pour les scénarios S2 et S3, c'est-à-dire si le nombre de produits par agglomérations présents dans le panier représentatif en grandes et moyennes surfaces est au moins multiplié par deux. En revanche, dans le cas du scénario S1 de stabilité du nombre de produits suivis par agglomération, on a un nombre insuffisant d'agglomérations dans les strates B et D dans l'échantillon d'agglomérations après concentration de la collecte.

Tableau 2 : Collecte après le passage aux données de caisses - comparaison entre les échantillons optimaux dans les trois scénarios étudiés et l'échantillon envisagé dans le cadre de la réorganisation de la collecte

Tranche d'agglomération	Collecte après passage aux données de caisses			
	S1 : Variance de première phase égale 0 sur le champ données de caisses	S2 : Variance de première phase égale 0 et variance de seconde phase divisée par deux sur le champ données de caisses	S3 :Variance égale à zéro sur le champ données de caisses	**Echantillon issu de la réorganisation de la collecte**
B (+ 100 000 habitants)	43	36	31	**37**
C (20 000 à 100 000 habitants)	14	12	10	**17**
D (2 000 à 20 000 habitants	22	18	16	**20**

Ainsi, dans le cadre de modélisation proposée par Ardilly et Guglielmetti et sur la base des calculs effectués sur les données 2009, l'échantillon d'agglomération envisagé suite la réorganisation de la collecte apparaît comme un échantillon admissible, c'est-à-dire permettant de conserver un niveau de précision pour la collecte globale au moins égal à celui observé avant le passage aux données de caisses, sous réserve de pouvoir multiplier par deux le nombre de produits suivis dans les grandes et moyennes surfaces, ce qui paraît un objectif raisonnable compte-tenu de la richesse des données de caisses.

Ces résultats sont cependant à prendre avec une grande prudence, compte-tenu de l'importance des hypothèses effectuées dans le cadre de la modélisation et également du fait que les calculs n'ont été menés que sur une seule année. Dès que possible, les calculs seront effectués sur les années 2003 à 2008 (et également sur l'année 2010 début 2011), afin de tester la robustesse des résultats présentés par rapport à l'année d'observation des prix. Il convient également de rappeler que l'échantillon optimal d'agglomérations peut également fluctuer fortement d'un recensement sur l'autre, et que de ce fait l'échantillon optimal qui sera calculé sur la base des nouvelles Unités Urbaines 2010 pourra différer de l'échantillon optimal calculé ici sur la base du RP 1990.

Annexe - Estimateurs de la variance de premier degré et de second degré pour le calcul de la précision de l'indice des prix

La modélisation utilisée ici (qui reprend celle d'Ardilly et Guglielmetti pour le calcul de la variance cible), est celle d'un tirage à deux degrés avec au premier degré un tirage d'agglomérations par sondage aléatoire simple au sein de l'ensemble des agglomérations d'une tranche d'agglomération donnée (la dimension ZEAT n'étant donc pas prise en compte).

La variance totale d'un indice var-agglo élémentaire s'écrit[12] :

$$V(\hat{I}(cc,v)) = (1-\frac{m(cc,v)}{M(cc)})\frac{S_{1P}^2(cc,v)}{m(cc,v)} + \frac{1}{m(cc,v)\cdot M(cc)}\sum_{i=1}^{M}\frac{S_{2P}^2(i,v)}{n(i,v)}$$

où $V_1(\hat{I}(cc,v)) = (1-\frac{m(cc,v)}{M(cc)})\frac{S_{1P}^2(cc,v)}{m(cc,v)}$ représente la variance de premier degré et $V_2(\hat{I}(cc,v)) = \frac{1}{m(cc,v)\cdot M(cc)}\sum_{i=1}^{M(cc)}\frac{S_{2P}^2(i,v)}{n(i,v)}$ la variance de second degré.

L'estimateur $\hat{V}(\hat{I}(v))$ de la variance totale obtenu en remplaçant les vraies dispersions sur l'ensemble de la population $S_{1P}^2(cc,v)$ et $S_{2P}^2(i,v)$ par les quantités observées sur l'échantillon $s_{1P}^2(cc,v)$ et $s_{2P}^2(i,v)$ est sans biais.

$$\hat{V}(\hat{I}(cc,v)) = (1-\frac{m(cc,v)}{M(cc)})\frac{s_{1P}^2(cc,v)}{m(cc,v)} + \frac{1}{m(cc,v)\cdot M(cc)}\sum_{i\in S_1}\frac{s_{2P}^2(i,v)}{n(i,v)}$$

En revanche, l'estimateur « naïf » de la variance de premier degré $\hat{V}_1^b(\hat{I}(cc,v)) = (1-\frac{m(cc,v)}{M(cc)})\frac{s_{1P}^2(cc,v)}{m(cc,v)}$ est biaisé et surestime la vraie valeur de la variance de premier degré. De même, l'estimateur « naïf » de la variance de second degré $\hat{V}_2^b(\hat{I}(cc,v)) = \frac{1}{m(cc,v)\cdot M(cc)}\sum_{i\in S_1}\frac{s_{2P}^2(i,v)}{n(i,v)}$ est biaisé et sous-estime la vraie valeur de la variance de second degré.

Les estimateurs sans biais des variances de premier et de second degré sont respectivement :

$$\hat{V}_1(\hat{I}(cc,v)) = (1-\frac{m(cc,v)}{M(cc)})(\frac{s_{1P}^2(cc,v)}{m(cc,v)} - \frac{1}{(m(cc,v))^2}\sum_{i\in S_1}\frac{s_{2P}^2(i,v)}{n(i,v)})$$

$$\hat{V}_2(\hat{I}(cc,v)) = \frac{1}{(m(cc,v))^2}\sum_{i\in S_1}\frac{s_{2P}^2(i,v)}{n(i,v)}$$

[12] En négligeant la correction de population finie au deuxième degré et en effectuant l'hypothèse que le nombre de magasins par agglomération est identique pour toutes les agglomérations d'une même tranche de taille d'agglomérations.

On notera ainsi que l'estimateur sans biais de la variance de premier degré peut prendre des valeurs négatives.

Sur la base d'une hypothèse d'indépendance du tirage des prix observés entre les différentes variétés[13], les variances à un niveau plus agrégé sont obtenues en faisant la somme pondérée - par le carré du poids - des variances des indices élémentaires.

Par exemple, pour l'indice des prix global $\hat{I} = \sum_{CC,\nu} w(cc,\nu)\hat{I}(cc,\nu)$, la variance totale vaut :

$$V(\hat{I}) = \sum_{CC,\nu} w^2(cc,\nu)V(\hat{I}(cc,\nu))$$

On a de même, $\hat{V}(\hat{I}) = \sum_{CC,\nu} w^2(cc,\nu)\hat{V}(\hat{I}(cc,\nu))$, $\hat{V}_1(\hat{I}) = \sum_{CC,\nu} w^2(cc,\nu)\hat{V}_1(\hat{I}(cc,\nu))$ et $\hat{V}_2(\hat{I}) = \sum_{CC,\nu} w^2(cc,\nu)\hat{V}_2(\hat{I}(cc,\nu))$

[13] Cette hypothèse est bien sûr discutable compte-tenu du fait que l'enquêteur effectue pour des raisons d'efficacité le plus grand nombre de relevés possibles dans un même point de vente. Néanmoins, l'existence d'un nombre de relevés maximal par variété et par point de vente (seuil dépendant du point de vente, fixé à un relevé pour les commerces traditionnels, deux pour les supermarchés et trois pour les hypermarchés) donne une plus grande légitimité à cette hypothèse.