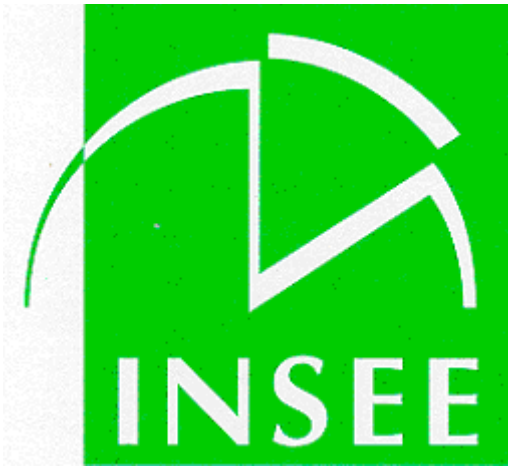
Direction des Statistiques Démographiques et Sociales

N° F1601

Échantillonnage des agglomérations de l'IPC pour la base 2015

Laurence JALUZOT et Patrick SILLARD

DOCUMENT DE TRAVAIL



Institut National de la Statistique et des Etudes Economiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES
Série des Documents de Travail
de la
DIRECTION DES STATISTIQUES DÉMOGRAPHIQUES ET SOCIALES

N°F1601

Échantillonnage des agglomérations de l'IPC pour la base 2015

LAURENCE JALUZOT ET PATRICK SILLARD

(DIVISION DES PRIX À LA CONSOMMATION)

Document de travail

janvier 2016

Échantillonnage des agglomérations de l'IPC pour la base 2015*

Laurence Jaluzot et Patrick Sillard†

Janvier 2016

Résumé

L'indice des prix à la consommation est un indice de Laspeyres chaîné annuellement. La base de l'indice désigne l'année sur laquelle, en moyenne, l'indice chaîné vaut 100. La notion de base d'indice recouvre aussi l'ensemble des conventions, en grande partie fixée par la réglementation européenne, qui prévalent dans la collecte et le calcul. Les changements de base sont en particulier requis lorsque ces conventions connaissent des modifications importantes. La base en vigueur jusqu'en 2015 date de 1998 mais le plan de sondage géographique n'avait pas été revu lors de son adoption et date donc du précédent changement de base (1990). En 2015, la base de l'indice des prix est revue, notamment dans le but de réviser l'échantillon des zones géographiques dans lesquelles les prix sont relevés.

Le présent document vise à exposer les principes d'échantillonnage retenus pour la sélection des produits suivis dans l'IPC en base 2015. Le plan de sondage retenu s'inspire des principes de 1990. Il a fait toutefois l'objet d'adaptations, notamment en vue d'atteindre deux objectifs : d'une part, conserver au maximum les zones de collecte existantes tout en maintenant le principe d'une sélection aléatoire des zones géographiques de collecte ; d'autre part, être en mesure de déterminer un nombre optimal de relevés à réaliser sur chaque zone géographique sélectionnée et pour chaque type de produits (variété) de sorte qu'à moyens de collecte fixés, la précision de l'IPC soit maximale.

En base 2015, les zones géographiques sélectionnées sont des unités urbaines. Un estimateur de l'indice des prix, fondé sur un sondage aléatoire des zones géographiques, est dérivé de l'expression exacte de l'indice que l'on cherche à estimer. Cet indice résulte de l'agrégation d'indices locaux déterminés à l'échelle de chaque unité urbaine, pour un type de produit donné. Puis ces indices élémentaires sont agrégés, chaque cellule élémentaire (croisement d'une unité urbaine et d'un type de produit) pesant dans l'agrégation en proportion de son poids dans la dépense de consommation des ménages. L'agrégation des échantillons d'indices élémentaires repose donc sur des poids qui tiennent compte, d'une part, de la probabilité d'inclusion de la zone géographique concernée dans l'échantillon et d'autre part, du poids de cette zone dans la dépense totale de consommation des ménages. Il est donc possible de dériver un estimateur sans biais de l'indice par un travail approprié sur les poids, de déterminer la variance de l'indice agrégé et d'optimiser cette dernière de façon à atteindre une précision optimale, sous contrainte de moyens. Ces différents éléments sont présentés dans ce document.

JEL Codes : E31 ; D01 ; C43 ; C83.

Mots-clés : indice de prix à la consommation, échantillonnage.

*Nous remercions Jérôme Accardo, Pascal Ardilly, Sébastien Faivre, Olivier Sautory ainsi que les rapporteurs du Comité du label pour les discussions fructueuses que nous avons eues avec eux dans le cadre de la préparation de cet article.

†Institut National de la Statistique et des Études Économiques – France.

Table des matières

Introduction

L'indice des prix à la consommation (IPC) est un indice de Laspeyres à panier fixe annuel, chaîné annuellement. En d'autres termes, l'indice est calculé, dans le courant de l'année, à partir de l'observation des prix des produits constituant le panier de biens et services suivis pour l'année considérée, en référence aux prix des mêmes produits observés au mois de décembre de l'année précédente. Puis cet indice-base décembre est chaîné pour constituer l'IPC proprement dit, en référence à l'année de base (actuellement 1998). L'Insee a décidé de réaliser, sur l'année 2015, un changement de base de l'IPC. Ce changement de base est justifié par une triple motivation : l'introduction de la nouvelle nomenclature européenne de diffusion des indices de prix à la consommation (COICOP), la rénovation du suivi des produits frais [1] destinée à aligner les concepts suivis sur les règlements européens et la rénovation de l'échantillon géographique du sondage IPC.

Le présent article vise à présenter les principes retenus pour la sélection des agglomérations de collecte de la base 2015, ainsi que la détermination, par optimisation de variance, du nombre de relevés à réaliser par type de produit et agglomération. Ce texte traite la question de la sélection des points de vente de métropole [2]. Il expose également, sur un plan pratique, les étapes mises en œuvre pour faire évoluer l'échantillon de la base 1998 vers celui de la base 2015. Un état des lieux de l'échantillon actuel (base 1998) est proposé en première section.

1 État des lieux : localisation des actuels points de vente de l'IPC

Les points de vente de l'IPC sont rattachés, en base 1998, à 96 agglomérations (au sens de l'IPC) métropolitaines. Au mois d'octobre 2013, il y avait, dans l'échantillon IPC, 26 104 points de vente actifs, c'est-à-dire des points de vente dans lesquels il existe un produit qui figure soit dans l'échantillon 2012, soit dans l'échantillon 2013 et qui ont déjà fait l'objet d'un relevé de prix en octobre 2012.
La notion d'agglomération IPC [3] n'est pas confondue avec celle des unités urbaines : elle est même protéiforme puisque dans certains cas, des unités urbaines disjointes appartiennent à la même agglomération IPC. À l'opposé, la collecte de Lyon ne concerne aujourd'hui qu'une petite partie de l'unité urbaine, de sorte que l'on peut considérer que l'agglomération IPC de Lyon est une fraction de l'unité urbaine de Lyon. Pourtant, dans l'esprit des concepteurs de l'échantillon IPC de 1990 (Ardilly & Guglielmetti 1992, Ardilly & Guglielmetti 1993), l'unité urbaine est l'individu statistique sélectionné dans le sondage géographique. Naturellement, cet individu était celui défini à l'époque du tirage, c'est-à-dire tel qu'il était défini dans le recensement de 1990. Quoiqu'il en soit, il est raisonnable de revenir à la notion d'unité urbaine qui est une notion précisément définie et pour laquelle on dispose d'un univers de sondage sur la base du dernier recensement de la population (2010).

La première étape d'identification de l'existant nécessite la localisation des points de vente actifs de métropole. Pour cela, on considère les 26 104 points de vente actifs de métropole en octobre 2013. 95,6% de ces points de

1. fruits, légumes, poissons, crustacés, fleurs et plantes.
2. Dans le cadre de la collecte IPC, des observations sont réalisées dans les DOMs. Du point de vue de l'échantillonnage, chaque DOM est considéré globalement comme un agrégat unique et la collecte est théoriquement réalisée sur l'ensemble du territoire. L'annexe E donne quelques éléments à propos du nombre de relevés, en relation avec la précision des indices calculés.
3. une autre façon de définir l'"agglomération IPC" est de considérer qu'elle correspond à l'agrégat géographique élémentaire, c'est-à-dire celui pour lequel on procède, à partir des prix des produits, à un calcul d'indice élémentaire. Au-delà de ce grain, l'agrégation est réalisée par agrégation de Laspeyres. Afin d'éviter l'ambiguïté du terme d'agglomération IPC, on parle d'*agglos* pour désigner un grain d'agrégat géographique élémentaire à partir de la section 2.

vente [4] comportent un numéro SIRET identifié, suite à l'opération de siretisation des points de vente conduite en 2011-2012 par les enquêteurs et les gestionnaires de sites-prix. Depuis lors, ce champ est tenu à jour lors des ouvertures/fermetures de points de vente.

Les points de vente qui comportent un numéro SIRET appariable avec la base SIRENE le sont dans une seconde étape. Ainsi, des 26 104 points de ventes actifs, 86,8% comportent un écho dans la base SIRENE. Ces points de vente sont immédiatement localisés, puisque la base SIRENE dispose pour chaque point de vente, de l'adresse et de l'identifiant de commune INSEE.
Pour les autres, on utilise un algorithme de reconnaissance de chaînes de caractères qui permet d'identifier la commune d'implantation du point de vente. Pour le reste (environ 400 points de vente), un codage manuel a été réalisé. Au final, l'ensemble des 26 104 points de vente actifs fin 2013 sont localisés à la commune. Il est dès lors possible de localiser la collecte de l'IPC et d'examiner sa répartition sur les unités urbaines métropolitaines. Cette analyse est réalisée par Jaluzot (2014). Elle est reproduite en annexe F.

2 Les principes de base de l'échantillon IPC

Faivre (2012) a proposé un schéma renouvelé de l'échantillon de la collecte IPC dans la perspective de l'introduction des données de caisses dans l'IPC, notamment en proposant une méthode de sondage à deux degrés stratifié fondé sur l'univers connu des données de caisses. Même si la base 2015 ne verra pas, *stricto sensu*, la substitution de données de caisses à de la collecte terrain, ces travaux permettent d'orienter la réflexion.

On limite ici le raisonnement au champ de la consommation localisée, c'est-à-dire celle pour laquelle une répartition géographique est connue et les prix varient spatialement. Le complément relève du champ des tarifs, c'est-à-dire de la consommation dont les prix ne varient pas spatialement ou pour laquelle on dispose *a priori* d'une information précise sur la variabilité spatiale des prix (exemple : le transport ferroviaire de voyageurs), ou celle pour laquelle la répartition spatiale est très concentrée par rapport à celle de la population générale (par exemple les achats de planches à voile). La définition précise du champ de la consommation localisée est partie intégrante de la définition des variétés. Préalablement à la définition de l'échantillon de l'IPC, il convient donc de séparer les variétés dont la logique de collecte correspond à une consommation localisée de celles dont on considère qu'elles correspondent à une consommation non localisée. Les prix correspondant à une consommation localisée sont en principe collectés sur le terrain par enquête, tandis que la consommation non localisée l'est de manière centralisée.

Pour le tirage de l'échantillon de produits dont les prix sont collectés sur le terrain (consommation localisée), les opérations suivantes doivent être réalisées :

1. Fixer les H strates parmi lesquelles on va sélectionner les agglos. Ces strates sont des croisements de zones géographiques (ZEAT [5] pour l'échantillon IPC base 1998) et de types d'agglomérations (pour

4. Certains points de vente, comme les photomaton, ne sont pas enregistrés en tant que point de vente dans la base SIRENE. Ils ne comportent donc pas d'identifiant SIRET. Afin de distinguer ces cas, le champ SIRET de la base des points de vente de l'IPC se voit attribuer une valeur fictive normalisée (14 caractère "9" successifs). Ces cas sont, dans cette statistique, considérés comme non siretisés. Il y 2,7% des points de vente actifs en octobre 2013 qui sont concernés par un SIRET fictif. Le complément (1,7%) comporte un champ SIRET vide.
5. Les ZEAT sont des regroupements de régions administratives (code entre parenthèses) : REGION PARISIENNE (1) - Ile de France ; BASSIN PARISIEN (2) - Bourgogne, Centre, Champagne-Ardenne, Basse et Haute Normandie, Picardie ; NORD (3) - Nord Pas-de-Calais ; EST (4) - Alsace, Franche-Comté, Lorraine ; OUEST (5) - Bretagne, Pays de la Loire, Poitou-Charentes ; SUD-OUEST (7) - Aquitaine, Limousin, Midi-Pyrénées ; CENTRE-EST (8) - Auvergne, Rhône-Alpes ; MEDITERRANEE (9) - Languedoc-Roussillon, Provence-Alpes-Côte d'Azur, Corse

l'échantillon IPC bases 1990 et 1998 – voir Ardilly & Guglielmetti (1992) et Faivre (2012)).

2. Déterminer le nombre n_h d'agglos retenues pour chaque strate h, $h \in \{1, \ldots, H\}$.
3. Pour une agglo a retenue, fixer le nombre de relevés $n_{a,v}$ à réaliser pour la variété v.

En première approximation, on souhaite que le nombre total de relevés $\sum_{v,a} n_{a,v}$ et que le nombre d'agglos $\sum_h n_h$ restent les mêmes qu'en base 1998. Dans un second temps, on cherchera à optimiser ces nombres selon un critère qui sera précisé par la suite.

Le plan de sondage est le suivant :

i. stratifié avec détermination du nombre n_h d'agglos pour chaque strate h ;

ii. au sein de chaque strate $h \in \{1, \ldots, H\}$, on procède à un sondage à deux degrés :

   a– on sélectionne n_h agglos selon un plan à préciser ;

   b– au sein de l'agglo a sélectionnée, on détermine un nombre de relevés $n_{a,v}$ à réaliser pour une variété v donnée et on sélectionne les relevés en question selon un plan à préciser.

3 Variable à estimer et estimateur

Notations : Dans toute cette partie, les indices se réfèrent à une variété donnée. L'indice de variété v est donc omis dans cette section.

L'indice d'ensemble est le fruit d'une agrégation de Laspeyres. En d'autres termes, quelle que soit la partition J de la consommation considérée, l'indice d'ensemble I, étant donnés les indices élémentaires I_j calculés pour les ensembles $j \in J$, vaut :

$$I = \sum_{j \in J} w_j I_j$$

où w_j est le poids de l'ensemble j dans la dépense de consommation des ménages. Par définition $\sum_{j \in J} w_j = 1$. Les indices élémentaires calculés dans l'IPC sont des indices calculés sur le couple (v, a). L'indice d'ensemble vaut donc, si on imagine une partition A du territoire en somme d'agglos (on omet le v de la variété),

$$I = \sum_{a \in A} w_a I_a \tag{1}$$

où w_a est le poids de l'agglo a dans la consommation des ménages (*i.e.* $\sum_a w_a = 1$).
Cet indice peut être estimé par sondage, d'une part en sélectionnant des agglos et d'autre part, en déterminant un estimateur de I_a à partir de l'observation de prix de produits appartenant à la variété d'intérêt v au sein de l'agglo a. Si la probabilité d'inclusion des produits contribuant à l'estimateur de I_a est indépendante de celle de l'agglo a dans l'échantillon (cas d'un sondage à deux degrés), alors la variance d'ensemble est la somme, d'une part, d'une variance intra spécifique à a et ne dépendant que du nombre de produits n_a sélectionnés dans l'échantillon sur lequel se fonde l'estimateur de I_a et d'autre part, d'une variance extra ne dépendant que des indices estimés $\hat{I}_a$ et du nombre n_a d'agglos sélectionnées. En effet, notons :

— $\mathbf{1}_{\mathscr{A}}(a)$ la variable de Cornfield caractérisant l'appartenance de l'agglo a à l'échantillon d'agglos sélectionnées $\mathscr{A}$, $\pi_a = \Pr(\mathbf{1}_{\mathscr{A}}(a) = 1)$ et $\pi_{a,a'} = \Pr\left(\{\mathbf{1}_{\mathscr{A}}(a) = 1\} \cap \{\mathbf{1}_{\mathscr{A}}(a') = 1\}\right)$;

— $\mathbf{1}_{\mathscr{S}_a}(i)$ la variable de Cornfield caractérisant l'appartenance du produit i à l'échantillon $\mathscr{S}_a$ de produits sélectionnés dans l'agglo a.

L'estimateur de l'indice d'ensemble I est estimé par $\hat{I}^{\mathscr{A}}$, fondé sur l'échantillon $\left(\mathscr{A}, \bigcup_{a\in\mathscr{A}} \mathscr{S}_a\right)$ et de la forme :

$$\hat{I}^{\mathscr{A}} = \sum_{a\in\mathscr{A}} \omega_a \hat{I}_a \tag{2}$$

où [6]

$$\left\{ \begin{array}{rcl} \hat{I}_a & = & \sum_{i\in\mathscr{S}_a} \omega_i \frac{p_i}{p_i^0} \\ \hat{I}_a & = & \prod_{i\in\mathscr{S}_a} \left(\frac{p_i}{p_i^0}\right)^{\omega_i} \end{array} \right. \quad \text{et} \quad \sum_{a\in\mathscr{A}} \omega_a = 1$$

Naturellement, pour que l'indice (2) soit sans biais, il est nécessaire en outre que $\mathbb{E}(\hat{I}_a) = I_a$.

Il est possible d'écrire $\hat{I}^{\mathscr{A}}$ à l'aide des variables de Cornfield :

$$\hat{I}^{\mathscr{A}} = \sum_{a\in A} \omega_a \hat{I}_a \times \mathbf{1}_{\mathscr{A}}(a) \tag{3}$$

Moyennant quoi, $\mathbb{E}(\hat{I}^{\mathscr{A}}) = I$, où I est donné par la formule (1) lorsque :

$$\omega_a = \frac{w_a}{\pi_a}$$

L'estimateur $\hat{I}^{\mathscr{A}}$ de I correspond donc à la formule :

$$\hat{I}^{\mathscr{A}} = \sum_{a\in\mathscr{A}} \frac{w_a \hat{I}_a}{\pi_a} \tag{4}$$

les poids $\frac{w_a}{\pi_a}$ étant de somme unitaire sur [7] $\mathscr{A}$.

La variance de $\hat{I}^{\mathscr{A}}$ s'écrit alors :

$$\mathrm{var}(\hat{I}^{\mathscr{A}}) = \sum_{a\in A} \left(\frac{w_a}{\pi_a}\right)^2 \mathrm{var}\left[\hat{I}_a \times \mathbf{1}_{\mathscr{A}}(a)\right] + \sum_{\substack{a\neq a' \\ (a,a')\in A^2}} \frac{w_a}{\pi_a}\frac{w_{a'}}{\pi_{a'}} \mathrm{cov}\left[\hat{I}_a \times \mathbf{1}_{\mathscr{A}}(a), \hat{I}_{a'} \times \mathbf{1}_{\mathscr{A}}(a')\right]$$

Les plans de sondage consistant à sélectionner $\mathscr{A}$ dans A et $\mathscr{S}_a$ dans l'ensemble des produits de l'agglo a étant indépendants, les variables $\mathbf{1}_{\mathscr{A}}(a)$ et $\hat{I}_a$ sont indépendantes. En revanche, les variables $\mathbf{1}_{\mathscr{A}}(a)$ et $\mathbf{1}_{\mathscr{A}}(a')$ sont, sauf hypothèse complémentaire, liées. Il découle de ce constat et de l'expression précédente que :

$$\begin{array}{rcl} \mathrm{var}(\hat{I}^{\mathscr{A}}) & = & \overbrace{\sum_{a\in A} \left(\frac{w_a}{\pi_a}\right)^2 \pi_a I_a^2 (1-\pi_a) + \sum_{\substack{a\neq a' \\ (a,a')\in A^2}} \frac{w_a}{\pi_a}\frac{w_{a'}}{\pi_{a'}} I_a I_{a'} (\pi_{aa'} - \pi_a \pi_{a'})}^{V1} \\ & & \underbrace{+ \sum_{a\in A} \left(\frac{w_a}{\pi_a}\right)^2 \pi_a \mathrm{var}(\hat{I}_a)}_{V2} \end{array} \tag{5}$$

On constate que la variance précédente se décompose en une somme de deux termes : le terme $V1$, appelé variance de premier degré, ne dépend que du tirage d'agglos, c'est la variance inter-agglos ; le terme $V2$, appelé variance de second degré, ne dépend que du tirage des produits dans les agglos sélectionnées, c'est la variance intra-agglos. Cette décomposition est une propriété classique des sondages à deux degrés (Tillé 2001, Ardilly

6. Le choix des ω_i n'est pas le cœur de cet article. En pratique, on suppose que les produits, au sein de la cellule (a, v), sont sélectionnés par sondage aléatoire simple avec remise (cf. annexe C), moyennant quoi, $\omega_i = 1/n_a$ où n_a est le nombre de produits observés pour la cellule (a, v) concernée.
7. et sont, le cas échéant, renormalisés pour ce faire.

2006). S'agissant de $V1$, tout se passe comme s'il s'agissait de la variance du π-estimateur du total (pour un échantillon $\mathscr{A}$ donné) :

$$\hat{y}^{\mathscr{A}} = \sum_{a \in \mathscr{A}} \frac{w_a I_a}{\pi_a} \tag{6}$$

où la caractéristique individuelle sommée est $y_a = w_a I_a$ (voir Tillé *op. cit.*, page 35). Ainsi, pour cette partie de l'estimateur $\hat{I}^{\mathscr{A}}$, on peut appliquer les résultats de théorie des sondages sur le π-estimateur du total en se rappelant que la variable sommée n'est pas I_a mais $w_a I_a$.

4 Optimisation de l'échantillon

Notations : À la différence du paragraphe 3, dans ce paragraphe, lorsqu'une quantité se réfère à une variété, ceci est explicitement indiqué en indice (v pour variété).

L'idée de base de l'optimisation consiste à déterminer les probabilités d'inclusion des agglos π_a et le nombre d'observations $n_{a,v}$, pour une variété donnée, de sorte que $\mathrm{var}(\hat{I}^{\mathscr{A}})$ donnée par l'expression (5) soit minimale. Naturellement, il faut procéder à quelques approximation pour pouvoir accéder à une expression de la variance qui soit utilisable analytiquement. Nous effectuons donc les hypothèses suivantes :

— **H1** – On procède à l'optimisation en deux étapes distinctes :

1. Les probabilités d'inclusion π_a des agglos sont déterminées *ex-ante* sur la base d'un critère global portant sur l'ensemble des variétés, comme proposé par Ardilly & Guglielmetti (1993).
2. Conditionnellement aux π_a de l'étape 1, on détermine pour la variété v, le nombre d'observations $n_{a,v}$ dans l'agglo a de façon à ce que le nombre total d'observations de la variété v soit n_v et que la variance, caractérisée par le facteur $V2$ dans l'expression (5) soit minimale. Le nombre d'observations $n_{a,v}$ est donc à ce stade conditionnel aux π_a et à n_v.

— **H2** – On suppose que le plan de sondage de l'échantillon d'observations pour une variété v dans une agglo a donnée est un sondage aléatoire simple avec remise de taille fixée égale à $n_{a,v}$. Cette hypothèse est acceptable dans la mesure où le taux de sondage est en général inconnu (on ne connaît pas le nombre de produits rattachés à une variété vendus dans une agglo donnée) et supposé très faible. Moyennant quoi, $\mathrm{var}(\hat{I}_{a,v}) = \frac{\sigma^2_{a,v}}{n_{a,v}}$, où $\sigma^2_{a,v}$ est un paramètre de variance [8] que l'on supposera indépendant de $n_{a,v}$.

— **H3** – w_v représente le poids de la variété v dans la dépense de consommation des ménages. $\sum_v w_v = 1$. $w_{a,v}$ correspond au poids de l'agglo a dans la dépense de consommation des ménages. Moyennant quoi, $\sum_{a \in A} w_{a,v} = w_v$, $\sum_v w_{a,v} = w_a$ où w_a est le poids de l'agglo a dans la dépense de consommation des ménages, et $w_{a,v}/w_v$ est le poids de l'agglo a dans la dépense de consommation des ménages de la variété v.

Sous les hypothèses **H1**, **H2** et **H3** précédentes, il est possible de déterminer le nombre d'observations requis pour le couple (a, v). En effet, la variance à minimiser de l'estimateur $\hat{I}_v$ de l'indice de la variété v, issu de l'agrégation des indices de var-agglo, est [9] :

$$\varphi(\mathbf{n}_v) = \frac{1}{w_v^2} \sum_{a \in A} \left(\frac{w_{a,v}}{\pi_a} \right)^2 \pi_a \frac{\sigma^2_{a,v}}{n_{a,v}} \tag{7}$$

8. Par exemple, pour une variété homogène, un estimateur de $\sigma^2_{a,v}$ est $\hat{\sigma}^2_{a,v} = \frac{1}{n_{a,v}-1} \sum_{i \in (a,v)} \left(\frac{p_i}{p_i^0} - \hat{I}_{a,v} \right)^2$ où $\hat{I}_{a,v} = \frac{1}{n_{a,v}} \sum_{i \in (a,v)} \frac{p_i}{p_i^0}$, p_i étant le prix courant du produit i et p_i^0 son prix au mois de base (voir annexe C pour plus de détails).
9. En effet, dans la formule (5), l'indice d'ensemble (i.e.pour une variété donnée) est la somme pondérée d'indices élémentaires d'agglos, les poids utilisés dans cette somme étant de somme unitaire. Par rapport aux notations introduites à l'hypothèse **H3**, il convient dès lors de considérer des poids normalisés, ce qui est le cas des $w_{a,v}/w_v$ dont la somme sur A est unitaire.

où $\mathbf{n}_v$ est le vecteur de composantes $(n_{a,v})_{a\in A}$. On peut donc la minimiser sous la condition que, en moyenne (i.e. en tenant compte de la probabilité de sélection de l'agglo a), le nombre d'observations pour la variété v soit égal à une quantité exogène n_v, c'est-à-dire que $\sum_{a\in A}\pi_a n_{a,v} = n_v$. Pour résumer, on définit le vecteur $\mathbf{n}_v = (n_{a,v})_{a\in A}$ comme [10] :

$$\mathbf{n}_v = \underset{\boldsymbol{\nu}}{\operatorname{argmin}} \left\{ \sum_{a\in A} \left(\frac{w_{a,v}}{\pi_a}\right)^2 \pi_a \frac{\sigma^2_{a,v}}{\nu_a} \,\middle|\, \sum_{a\in A} \pi_a \nu_a = n_v \right\} \tag{8}$$

Finalement, on obtient [11] :

$$n_{a,v} = \frac{w_{a,v}\sigma_{a,v}/\pi_a}{\sum_{a\in A} w_{a,v}\sigma_{a,v}} \times n_v \tag{9}$$

Ce résultat conduit à un nombre d'observations par couple (a, v) inconditionnel, c'est-à-dire qu'il ne dépend pas du choix de l'échantillon $\mathscr{A}$. En contrepartie, le nombre $n_{a,v}$ est aléatoire. On pourrait à l'inverse calculer un nombre optimal $n^{\mathscr{A}}_{a,v}$ conditionnellement à l'échantillon $\mathscr{A}$. Dans ce cas, $n^{\mathscr{A}}_{a,v}$ est défini par [12] :

$$\mathbf{n}^{\mathscr{A}}_v = \underset{\boldsymbol{\nu}}{\operatorname{argmin}} \left\{ \sum_{a\in \mathscr{A}} \left(\frac{w_{a,v}}{\pi_a}\right)^2 \frac{\sigma^2_{a,v}}{\nu_a} \,\middle|\, \sum_{a\in \mathscr{A}} \nu_a = n_v \right\} \tag{10}$$

L'intérêt de cette seconde approche est qu'il n'y a plus d'aléa sur le nombre d'observations par couple (a, v) dès que $\mathscr{A}$ est sélectionné : le nombre total d'observations est fixé et égal à la somme des observations réalisées sur l'ensemble des agglos de l'échantillon $\mathscr{A}$.
On montre [13] que :

$$n^{\mathscr{A}}_{a,v} = \frac{w_{a,v}\sigma_{a,v}/\pi_a}{\sum_{a\in \mathscr{A}} w_{a,v}\sigma_{a,v}/\pi_a} \times n_v \tag{11}$$

Les deux estimateurs (9) et (11) sont en tout point comparables. En effet, dans le premier, la présence de π_a au numérateur permet de dilater le poids de l'agglo a en tenant compte du fait que si elle est sélectionnée, elle "représente" davantage qu'elle-même (car $\pi_a \leqslant 1$) ; au dénominateur la somme des poids est réalisée sur tout l'univers, donc la probabilité d'inclusion n'apparaît pas. À l'inverse, pour le second, la somme au dénominateur étant réalisée sur l'échantillon sélectionné $\mathscr{A}$, il convient là-aussi de dilater les poids des agglos sélectionnées pour tenir compte du fait qu'elles représentent davantage qu'elles-mêmes. En espérance, les deux estimateurs sont cohérents.

À ce stade, il nous reste à préciser les quantités π_a (et en particulier le nombre d'agglos retenues) et le nombre n_v d'observations pour chaque variété. Pour préciser le nombre total d'observations $n = \sum_v n_v$, nous effectuons les hypothèses suivantes, complémentaires aux précédentes :

— **H4** – le nombre d'observations total (pour le champ de la consommation localisée) résulte d'une optimisation de la variance de second degré sous contrainte de coût. Le coût d'ensemble est C et le coût d'un relevé élémentaire pour la variété v est c_v. Ces paramètres sont exogènes. Les indices de variété obtenus par agrégation des indices pour les couples (a, v) sont indépendants. Le nombre d'observations effectué pour le couple (a, v) est celui résultant de l'optimisation précédente, et donné par la relation (9).

10. Le facteur $1/w_v^2$ est omis à ce stade du programme d'optimisation car sa présence ne modifie pas la solution.
11. voir annexe B avec $a_i \to w_{a,v}\sigma_{a,v}/\sqrt{\pi_a}$, $\alpha_i \to \sqrt{\pi_a}$ et $K \to n_v$
12. Voir aussi note N°10.
13. voir annexe B avec $a_i \to w_{a,v}\sigma_{a,v}/\pi_a$, $\alpha_i \to 1$ et $K \to n_v$

On note c_v le coût élémentaire d'un relevé de la variété v, C le coût total de l'opération de collecte. L'indice d'ensemble résulte d'une agrégation des indices élémentaires de variétés I_v conformément à la relation [14] :

$$I = \sum_v w_v I_v$$

où w_v est le poids de la variété v dans la dépense de consommation des ménages. Il en découle que la variance de second degré pour l'indice d'ensemble vaut [15] ($\widehat{V2}^{\mathscr{A}} = \mathrm{var}^{2d}(\hat{I}^{\mathscr{A}})$) :

$$\widehat{V2}^{\mathscr{A}} = \sum_v w_v^2 \times \left\{ \frac{1}{w_v^2} \sum_{a \in \mathscr{A}} \left(\frac{w_{av}}{\pi_a} \right)^2 \frac{1}{n_v} \sigma_{a,v}^2 \left[\frac{w_{av}\sigma_{av}/\pi_a}{\sum_{a \in \mathscr{A}} w_{av}\sigma_{av}/\pi_a} \right]^{-1} \right\}$$

Après simplification, cette variance s'écrit :

$$\widehat{V2}^{\mathscr{A}} = \sum_v \frac{(\kappa_v^{\mathscr{A}})^2}{n_v} \tag{12}$$

où

$$\kappa_v^{\mathscr{A}} = \sum_{a \in \mathscr{A}} w_{av}\sigma_{av}/\pi_a$$

On détermine le vecteur $\mathbf{n} = (n_1, \ldots, n_V)$ en minimisant la variance sous contrainte de coût global de collecte :

$$\mathbf{n} = \underset{\boldsymbol{\nu}}{\mathrm{argmin}} \left\{ \sum_v \frac{(\kappa_v^{\mathscr{A}})^2}{\nu_v} \middle| \sum_v c_v \nu_v = C \right\} \tag{13}$$

On montre [16], pour tout $v \in \{1, \ldots, V\}$, que l'estimateur du nombre total d'observations par variété v, conditionnel à l'échantillon d'agglos sélectionnées $\mathscr{A}$ est :

$$n_v^{\mathscr{A}} = \frac{C}{c_v} \times \frac{\kappa_v^{\mathscr{A}} \sqrt{c_v}}{\sum_v \kappa_v^{\mathscr{A}} \sqrt{c_v}} \tag{14}$$

Cette expression correspond à une allocation de Neyman (Ardilly 2006).
On en déduit que la variance totale optimisée vaut :

$$\underline{\widehat{V2}}^{\mathscr{A}} = \frac{1}{C} \left(\sum_v \kappa_v^{\mathscr{A}} \sqrt{c_v} \right)^2 \tag{15}$$

C'est la borne inférieure de la variance que l'on peut obtenir avec l'échantillon $\mathscr{A}$ et pour un coût de collecte C.

5 Les formules de variance des micro-indices

Ces formules ont été examinées par Ardilly & Guglielmetti (1992) dans un contexte où les formules élémentaires étaient légèrement différentes de celles pratiquées aujourd'hui. En particulier, à l'époque, la formule appliquée pour les variétés hétérogènes consistait en une moyenne arithmétique de rapports de prix ; la formule pratiquée aujourd'hui est une moyenne géométrique de rapports de prix. Les formules ont été revues par Petit (2014) et Balcone (2014). Les formules proposées ici sont démontrées en annexe C. Elles correspondent aux formules établies pour la variance intra au niveau des micro-indices par Petit (2014) et Balcone (2014).

14. Pour être complet, avec ces notations, nous avons en particulier $\sum_v w_v = 1$ et $\sum_{a \in A} w_{a,v} = w_v$. Même si dans le cas général, $w_{a,v}$ est quelconque, il pourra, par exemple par secteur de variété ou strate d'agglos $\mathcal{S}(a,v)$, être compris comme le produit de poids marginaux : $w_{a,v} = w_a^{\mathcal{S}(a,v)} \times w_v$. Moyennant quoi, on a alors $\sum_{a \in A} w_a^{\mathcal{S}(a,v)} = 1$.
15. On repart ici de la formule (7) et on somme en pondérant chaque indice de variété $\hat{I}_v$ par son poids w_v (mis au carré dans la formule de variance). On s'appuie ici sur l'expression de l'estimation de la variance de $\hat{I}_v$ fondée sur l'échantillon $\mathscr{A}$, soit les relations (10) et (11).
16. voir annexe B avec $a_i \to \kappa_v^{\mathscr{A}}$, $\alpha_i \to \sqrt{c_v}$ et $K \to C$

5.1 Variétés homogènes

Un indice de prix d'agglo pour une variété homogène donnée (on omet ici l'indice de variété et celui d'agglo) est calculé à l'aide d'une formule d'indice de Laspeyres : soient $(p_i^t)_{i\in\{1,\dots n\}}$ les prix des n biens composant le panier d'intérêt à l'instant t et $(p_i^0)_{i\in\{1,\dots n\}}$ les prix de ces mêmes biens à l'instant de base, alors l'indice estimé est [17] :

$$\hat{I} = \frac{\frac{1}{n}\sum_{i=1}^{n} p_i^t}{\frac{1}{n}\sum_{i=1}^{n} p_i^0}$$

Sous l'hypothèse où les n biens ont été sélectionnés par sondage aléatoire simple sans remise dont le taux de sondage est négligeable, alors on montre (voir annexe C) qu'un estimateur de la variance (intra) est :

$$\widehat{\mathrm{var}(\hat{I})} = \left(\frac{1}{\hat{\bar{P}}^0}\right)^2 \frac{1}{n(n-1)} \sum_{i\in\mathscr{S}} \left(p_i^t - \hat{I} p_i^0\right)^2 \tag{16}$$

où $\hat{\bar{P}}^0$ est la moyenne empirique des prix de base sur l'échantillon de produits retenus.
Par rapport au paramètre σ_{av} évoqué au paragraphe 4 (hypothèse **H2**), il suffit de considérer l'expression précédente multipliée par le nombre d'observations qui concourent à la somme [18] et d'en prendre la racine carrée.

5.2 Variétés hétérogènes

Un indice de prix d'agglo pour une variété hétérogène donnée (on omet ici l'indice de variété et celui d'agglo) est calculé à l'aide d'une formule d'indice de Laspeyres géométrique. Avec les mêmes notations qu'au paragraphe précédent, l'indice estimé est (n biens composent le panier) :

$$\hat{I} = \left(\prod_{i=1}^{n} \frac{p_i^t}{p_i^0}\right)^{1/n}$$

Sous l'hypothèse où les n biens ont été sélectionnés par sondage aléatoire simple sans remise dont le taux de sondage est négligeable, alors on montre (voir annexe C) qu'un estimateur de la variance (intra) est :

$$\widehat{\mathrm{var}(\hat{I})} = \frac{\hat{I}^2}{n(n-1)} \sum_{i\in\mathscr{S}} \left[\ln\left(\frac{p_i^t}{p_i^0}\right) - \ln \hat{I}\right]^2 \tag{17}$$

Comme précédemment, s'agissant du paramètre σ_{av} évoqué au paragraphe 4 (hypothèse **H2**), il suffit de considérer l'expression précédente multipliée par le nombre d'observations qui concourent à la somme et d'en prendre la racine carrée.

5.3 Variétés produits frais

Le calcul des indices de produits frais répond à une mécanique assez différente de celle utilisée pour les variétés ordinaires, puisque l'agrégation est réalisée au niveau poste-strate, et non au niveau var-agglo. Par ailleurs, les produits frais pèsent peu dans l'échantillon et sont observés dans la plupart des agglos IPC. L'enjeu à les inclure dans l'analyse, notamment pour la sélection des agglos, est donc relativement faible tandis que leur

17. Dans tout ce texte, on omet l'éventuel facteur multiplicatif 100 qui permet, en l'absence de variation de prix, de fixer le niveau de base à 100. En l'absence de ce facteur, le niveau de base est égal à 1.

18. En effet, on définit $\sigma_{a,v}$ de sorte que $\mathrm{var}(\hat{I}_{a,v}) = \frac{\sigma_{a,v}^2}{n_{a,v}}$, où $n_{a,v}$ est le nombre d'observations de l'échantillon pour la variété v dans l'agglo a.

intégration contribuerait à accroître la complexité de l'algorithme. Comme Ardilly & Guglielmetti (1992), nous ne prenons pas en compte les produits frais dans cette analyse.

6 Le plan du sondage de premier degré

Comme évoqué ci-dessus (paragraphe 3), la variance du sondage de premier degré s'apparente à celle d'un π-estimateur du total de la variable d'agglo $w_{a,v}I_{a,v}$ où $w_{a,v}$ est le poids du couple (a, v) dans la dépense de consommation des ménages. Il est clair que cette variable est corrélée à toute variable de taille d'agglo [19].

En conséquence, il serait justifié (voir Tillé (2001)) d'adopter un plan de sondage des agglos avec probabilités d'inclusion inégales proportionnelles à une variable de taille. En pratique, ce constat donne un objectif à suivre pour le choix des probabilités d'inclusion des agglos dans l'échantillon. Un tel plan de sondage permet d'assurer une précision optimisée du π-estimateur (Tillé 2001) par rapport, par exemple, à un sondage à probabilités d'inclusion égales. Cependant, le choix des agglos est aussi dicté par l'échantillon antérieur d'agglos. Donc le plan de sondage ne peut pas suivre *stricto-sensu* un plan qui serait défini de manière purement *ad hoc*, indépendamment de l'échantillon actuel d'agglos.

Le sondage destiné à la sélection des agglos est un sondage stratifié par zone géographique (ZG) croisées avec les classes de tailles d'agglos (TA). Ces strates définissent une partition de l'espace géographique. On note les strates $h \in \{1, \ldots, H\}$.

Pour cet exercice, on distingue deux étapes :

1. détermination du nombre n_h d'agglos par strate en minimisant un critère une fonction de variance [20] d'ensemble (de premier degré) sous contrainte de coût ;
2. sondage au sein de chaque strate avec probabilité d'inclusion approximativement proportionnelle à une variable de taille.

Ces deux étapes sont précisées ci-après.

6.1 Le nombre d'agglos par strate

Le nombre d'agglos par strate doit être déterminé de sorte que la variance de premier niveau du sondage à deux degrés soit minimale. Au sein de chaque strate, les agglos sont sélectionnées par sondage aléatoire avec probabilité d'inclusion proportionnelle à une variable de taille. On notera π_a la probabilité d'inclusion dans l'échantillon d'une agglo a. Elle appartient à une strate h donnée. Dans la mesure où le nombre d'agglos par strate, noté n_h pour la strate h, est la variable d'intérêt à ce stade, il convient de déterminer ce nombre de sorte que la variance d'ensemble soit minimale, sous contrainte d'un coût lié au nombre d'agglos retenu pour chaque strate. Dans les étapes exposées au paragraphe 4, la liste des agglos est fixée et les coûts appliqués pour l'optimisation sont ceux de la collecte de relevés, conditionnellement à la liste des agglos et au nombre total de relevés à réaliser sur l'ensemble des agglos. Les coûts qui doivent être pris en compte dans l'étape présente, qui consistent à ventiler un nombre total de relevés fixé sur des agglos en plus ou moins grand nombre, s'apparentent à des coûts d'approche, pour les enquêtes ménages. Faute d'information particulière,

19. De ce point de vue, différentes variables sont susceptible de correspondre à $w_{a,v}$ et le choix, sur le plan des principes du sondage, n'est pas essentiel pourvu que la variable soit bien liée à la taille des agglos : on peut penser à une variable de taille liée à la démographie ou à une variable prenant en compte la géographie de la consommation des ménages. Des éléments de cadrage sont proposés par Jaluzot (2014) – voir annexe F.
20. On parle ici de fonction de variance d'ensemble car celle-ci est une forme approximée de ce que serait la variance d'ensemble si un tirage sans remise avec probabilité d'inclusion proportionnelle à une variable de taille était réalisé.

il n'y a pas lieu de considérer que ces coûts diffèrent d'une agglo à l'autre. Nous supposerons donc que la fonction de coût est simplement la somme des nombres d'agglos obtenus sur chacune des strates (soit un coût unitaire par agglo incluse dans l'échantillon). Cette hypothèse est également celle retenue par Ardilly & Guglielmetti (1992) et Faivre (2012) dans leurs travaux.

Comme vu à la relation [21] (6), on estime la variance du π-estimateur du total de la variable $w_{a,v}I_{a,v}$ sur la strate. Ardilly (2014) propose d'utiliser, pour le calcul de variance, une formule découlant de l'approximation proposée par Ardilly (2006) – page 156 :

$$V1_{h,v} = \sum_{a\in h} \pi_{a,h}(1-\pi_{a,h}) \left(\frac{w_{a,v}I_{a,v}}{\pi_{a,h}} - \frac{1}{N_h} \sum_{j\in h} \frac{w_{j,v}I_{j,v}}{\pi_{j,h}} \right)^2$$

où N_h est le nombre total d'agglos dans la strate h. On suppose donc que $\pi_{a,h} = w_a \times n_h/W_h$ où n_h est le nombre (inconnu) d'agglos à sélectionner dans la strate h, w_a est défini conformément à la relation (1) et $W_h = \sum_{a\in h} w_a$. Puis, comme on ne connaît pas le poids en dépense de consommation sur une maille élémentaire (a, v) mais que les poids marginaux sont connus, nous posons que [22] $w_{a,v} = w_a \times w_v$. Il en découle que :

$$V1_{h,v}(n_h) = w_v^2 W_h \sum_{a\in h} w_a \left(\frac{1}{n_h} - \frac{w_a}{W_h}\right) \left(I_{a,v} - \frac{1}{N_h} \sum_{j\in h} I_{j,v} \right)^2 \tag{18}$$

Formellement, il convient de retenir le vecteur $\mathbf{n} = (n_1, \ldots, n_H)$ comme solution du problème de minimisation de la variance sous contrainte que le nombre total d'agglos soit égal à $\mathcal{N}$:

$$\mathbf{n} = \underset{\boldsymbol{\nu}}{\operatorname{argmin}} \left\{ \sum_{h=1}^{H} \sum_{v} V1_{h,v}(\nu_h) \,\middle|\, \sum_{h=1}^{H} \nu_h = \mathcal{N} \right\}$$

Compte tenu du très faible nombre d'agglos par strate figurant dans l'échantillon actuel de l'IPC, il n'est pas possible d'estimer convenablement une variance d'observation des indices variétés-agglomérations $\left(I_{a,v} - \frac{1}{N_h}\sum_{j\in h} I_{j,v}\right)^2$. Par conséquent, nous choisissons de considérer que la variance précédente est identique pour toutes les strates et est donc égale à un paramètre σ_v^2 ne dépendant pas de h. Finalement, le problème d'optimisation revient à [23] :

$$\mathbf{n} = \underset{\boldsymbol{\nu}}{\operatorname{argmin}} \left\{ \sum_{h=1}^{H} \frac{W_h^2}{\nu_h} \,\middle|\, \sum_{h=1}^{H} \nu_h = \mathcal{N} \right\} \tag{19}$$

Finalement [24] (voir annexe B),

$$n_h = \mathcal{N} \times W_h \tag{20}$$

21. Dans la somme correspondante, A désigne l'ensemble des agglos du territoire. En d'autres termes, $A = \bigcup_{\{1,\ldots,H\}} h$.
22. Avec ces notations $\sum_a w_a = 1$, $\sum_v w_v = 1$ et $\sum_{a,v} w_{a,v} = 1$.
23. Les multiplicateurs positifs et autres constantes qui s'ajoutent à la fonction objectif sont simplifiés dans cette expression.
24. Par rapport aux travaux antérieurs, ce développement appelle la remarque suivante. Le nombre d'agglos enquêtées par strate est ici le fruit d'une minimisation de variance sous la contrainte que la somme des coûts marginaux de collecte d'une agglo soient les mêmes qu'actuellement. Cette approche est différente de celle proposée par Ardilly & Guglielmetti (1992) et reprise par Faivre (2012) qui minimisent la somme des coûts (avec une hypothèse de coûts marginaux par strate unitaire) sous la contrainte que la variance est la même qu'auparavant. La forme proposée ici est plus simple à mettre en œuvre et théoriquement duale, sous la réserve que le coût et la variance de référence des relations de contrainte associées aux deux problèmes soient déduits l'un de l'autre, ce qui n'est vraisemblablement pas le cas. Pour le calcul d'optimisation, voir annexe B avec $a_i \to W_h$, $\alpha_i \to 1$ et $K \to \mathcal{N}$ où $\sum_{h=1}^{H} W_h = 1$.

6.2 La sélection dans chaque strate conditionnellement à $\mathbf{n_h}$

Comme on l'a vu, il est judicieux d'adopter un plan de sondage tel que les probabilités d'inclusion s'approchent d'une proportionnalité à une variable de taille d'agglo. On peut faire l'hypothèse que dans l'échantillon actuel, c'est bien le cas. En effet, cette propriété est mentionnée par Ardilly & Guglielmetti (1992) – page 75 : *"La taille d'échantillon par groupe*[25] *est à peu près proportionnelle à la taille moyenne des agglos du groupe auquel elles appartiennent."*

La sélection des agglos s'opère conditionnellement à leur appartenance à l'échantillon d'agglo de l'IPC-base 1998. Nous notons $\mathscr{I}$ l'échantillon IPC-base 1998 et $\mathscr{X}$ l'échantillon nouvelle base. On raisonne ci-après sur des ensembles d'agglos sur lesquels un sondage aléatoire est réalisé.

On cherche à faire en sorte que la probabilité d'inclusion d'une agglo a dans l'échantillon $\mathscr{X}$ soit égale à une valeur $\pi_a^{\mathscr{X}}$ connue, tout en faisant dépendre la probabilité de sélection dans l'échantillon $\mathscr{X}$ de l'appartenance de l'agglo à l'échantillon $\mathscr{I}$. En effet, l'idée est de faire en sorte, autant que possible, qu'une agglo présente dans l'échantillon $\mathscr{I}$ ait un maximum de chance de figurer dans $\mathscr{X}$ et, en particulier, davantage qu'une agglo qui ne figure pas dans $\mathscr{I}$.

On a classiquement :

$$\Pr(a \in \mathscr{X}) = \Pr(a \in \mathscr{X} | a \in \mathscr{I}) \times \Pr(a \in \mathscr{I}) + \Pr(a \in \mathscr{X} | a \notin \mathscr{I}) \times \Pr(a \notin \mathscr{I})$$

On fait l'hypothèse ici qu'on connaît $\Pr(a \in \mathscr{I}) = \pi_a^{\mathscr{I}}$ et $\Pr(a \notin \mathscr{I}) = 1 - \pi_a^{\mathscr{I}}$. Moyennant quoi,

$$\pi_a^{\mathscr{X}} = \pi_a^{\mathscr{I}} \times \Pr(a \in \mathscr{X} | a \in \mathscr{I}) + (1 - \pi_a^{\mathscr{I}}) \times \Pr(a \in \mathscr{X} | a \notin \mathscr{I}) \tag{21}$$

En pratique, on souhaite évoluer le moins possible par rapport à l'échantillon d'origine. En d'autres termes, on cherche à maximiser la probabilité d'inclusion dans le nouvel échantillon d'une agglo qui figurerait dans l'ancien (i.e.la valeur de $\Pr(a \in \mathscr{X} | a \in \mathscr{I})$) tandis que la relation (21) serait vérifiée.

On montre (cf. annexe D) que deux situations peuvent survenir :

— soit $\pi_a^{\mathscr{X}} \leqslant \pi_a^{\mathscr{I}}$ et dans ce cas il convient de retenir les probabilités conditionnelles d'inclusion suivantes :

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

— soit $\pi_a^{\mathscr{X}} > \pi_a^{\mathscr{I}}$ et dans ce cas il convient de retenir les probabilités conditionnelles d'inclusion suivantes :

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

En faisant l'hypothèse, pour un groupe[26] G d'agglos sur lequel on souhaite sélectionner un échantillon aléatoire d'agglos, que la probabilité d'inclusion est proportionnelle à une variable T_a de taille d'agglo, alors

$$\pi_a^G = n_G \times \frac{T_a}{\sum_{a \in G} T_a}$$

où n_G est le nombre d'agglos sélectionnées dans le groupe G (ce nombre est exogène à ce stade de la procédure). On peut construire de la sorte les $\pi_a^{\mathscr{I}}$ et les $\pi_a^{\mathscr{X}}$. Puis, sur G on dispose d'une liste $G^{\mathscr{I}}$ d'agglos appartenant à l'échantillon $\mathscr{I}$. Conformément à ce qui précède, sur G, on applique l'algorithme de sélection suivant :

25. NDR : comprendre strate.
26. Typiquement, une strate.

1. Si $a \in G^{\mathscr{I}}$, deux cas peuvent se produire :
   (a) si $\pi_a^{\mathscr{X}} \leqslant \pi_a^{\mathscr{I}}$, alors a est retenue dans $\mathscr{X}$ avec une probabilité $\pi_a^{\mathscr{X}} / \pi_a^{\mathscr{I}}$;
   (b) si $\pi_a^{\mathscr{X}} > \pi_a^{\mathscr{I}}$, alors a est retenue dans $\mathscr{X}$.
2. Si $a \notin G^{\mathscr{I}}$, deux cas peuvent se produire :
   (a) si $\pi_a^{\mathscr{X}} \leqslant \pi_a^{\mathscr{I}}$, alors a n'est pas retenue dans $\mathscr{X}$;
   (b) si $\pi_a^{\mathscr{X}} > \pi_a^{\mathscr{I}}$, alors a est retenue dans $\mathscr{X}$ avec une probabilité $(\pi_a^{\mathscr{X}} - \pi_a^{\mathscr{I}})/(1 - \pi_a^{\mathscr{I}})$.

Considérons un groupe G dans lequel on doit sélectionner n_G agglos. D'un point de vue pratique, dans le cas 1b précédent, les agglos sont retenues. On suppose qu'il y en a $n_G^{(1)}$ dans ce cas. Puis, pour les autres agglos, on procède par tirage systématique proportionnel à la variable de probabilité d'inclusion conditionnelle indiquée aux 1a et 2b précédents en sélectionnant $n_G - n_G^{(1)}$ agglos. Cette sélection est implémentée sous SAS comme indiqué par Ardilly (2006) – page 157.

7 Application numérique

7.1 Détermination du nombre $\mathbf{n_h}$

7.1.1 Choix des strates

Dans un premier temps, on fixe la nature des strates retenues. Pour l'échantillon IPC, on procède comme Ardilly & Guglielmetti (1992) en réunissant les agglos par taille d'unité urbaine [27] et en divisant le territoire métropolitain en ZISP [28] (zone d'intervention d'un site prix).

7.1.2 Choix des poids $\mathbf{w_{a,v}}$

Comme vu au paragraphe 6, le poids $w_{a,v}$ se décompose en un produit $w_a \times w_v$. Le poids des variétés est celui issu des chiffres de la Comptabilité nationale étant donnée la liste des variétés à la fin de l'année 2013. S'agissant des poids d'agglos w_a, on utilise une proxy de la dépense de consommation des ménages au lieu d'achat. Cette variable, construite par Jaluzot (2014), est fondée sur [29] les éléments suivants :

1. La distribution de la dépense de consommation alimentaire des ménages selon la taille d'agglo est connue par l'enquête Budget des familles 2011. La variable de taille d'unité urbaine utilisée ici est celle des données du recensement (2010) codée en 9 catégories [30].
2. Les poids précédents sont ensuite ventilés, au sein des classes de taille d'unités urbaines, sur l'ensemble des unités urbaines de la classe, proportionnellement à la population municipale 2010.

27. On distingue 5 classes : A=UU de Paris ; B2=UU de plus de 200 000 habitants ; B1=UU de 100 000 à 200 000 habitants ; C=UU de 20 000 à 100 000 habitants ; D=UU de 2 000 à 20 000 habitants. Les communes rurales sont dans une classe H complémentaire.
28. Une zone d'intervention des sites prix (ZISP) est une partition de l'espace métropolitain en 7 zones. C'est une somme de régions administratives, à l'exception du département de l'Yonne rattaché à la ZISP du site-prix de Saint-Quentin-en-Yvelines et non à celle de Lyon et du département des Ardennes rattaché à la ZISP du site-prix de Lille et non à celle de Nancy.
29. Voir annexe F.
30. Code "taille de l'unité urbaine" (basé sur la population municipale au recensement 2007 pour les UU 2010) : 0-Rural ; 1-Unités urbaines de 2 000 à 4 999 habitants ; 2-Unités urbaines de 5 000 à 9 999 habitants ; 3-Unités urbaines de 10 000 à 19 999 habitants ; 4-Unités urbaines de 20 000 à 49 999 habitants ; 5-Unités urbaines de 50 000 à 99 999 habitants ; 6-Unités urbaines de 100 000 à 199 999 habitants ; 7-Unités urbaines de 200 000 à 1 999 999 habitants ; 8-Agglomération de Paris.

7.1.3 Le calcul de n_h

Avec les éléments précédents, on applique la formule (20). Pour $\mathcal{N}$, on prend le nombre total d'agglos devant figurer dans l'échantillon (fixé de manière exogène à 100 pour la métropole [31]). Il arrive que pour une strate donnée h, la formule aboutisse à un nombre d'agglos sélectionnées supérieur à celui qui figure dans la strate. Dans ce cas, il convient de saturer le nombre d'agglos sélectionnées pour la strate au nombre disponible : $n_h = N_h$. Puis de relancer l'algorithme de calcul en réduisant l'échantillon à sélectionner à un échantillon de taille $\mathcal{N} - N_h$. Les poids W_h doivent être renormalisés sur le nouvel univers du sondage (l'ensemble d'origine moins la strate saturée h) de façon à être de somme unitaire sur le nouvel univers. On procède de la même façon itérativement jusqu'à ce que tous les n_h soient déterminés.

7.2 La sélection des agglomérations

Pour chaque strate, nous procédons conformément à l'algorithme développé au paragraphe 6.2. Nous raisonnons, pour les agglos, sur l'enveloppe des unités urbaines au sens du RP de 2010. Nous supposons que la probabilité d'inclusion d'une telle agglo dans l'échantillon de l'IPC-base 1998 est proportionnelle au nombre d'habitants compris dans cette agglo au sens de la population sans double compte de 1990. Le coefficient de proportionnalité est tel que la somme des probabilités d'inclusion sur une strate soit égale au nombre d'unités urbaines couvertes par la collecte IPC actuelle au seuil de 100 relevés minimum par mois [32]. Les agglos sont sélectionnées de sorte que leur probabilité d'inclusion dans l'échantillon-base 2015 soit proportionnelle à une variable de pondération, fondée sur une proxy de la dépense de consommation des ménages localisée au lieu d'achat [33].

Le tableau 1 donne le nombre d'agglos par strate selon la variable de pondération utilisée, conditionnellement au fait que l'échantillon sélectionné comprend 100 unités urbaines. La colonne n_h du tableau indique le nombre d'unité urbaines sélectionnées dans le tirage systématique.

7.3 Détermination du nombre de relevés par variété et par agglomération

Pour le calcul du nombre de relevés n_v par variété, on utilise pour les coûts c_v les temps utilisés pour calculer les temps de travail des enquêteurs (Sillard 2012a, Sillard 2012b) et reportés dans la table 3.
À l'aide de ces éléments et des probabilités d'inclusion qui découlent du mode de sélection des agglos décrit au paragraphe 7.2, on en déduit le nombre de relevés requis par variété. Enfin, en appliquant la formule (11), on en déduit le nombre d'observations par variété et agglo.

L'optimisation du nombre de relevés par unité urbaine et par variété conduit à réviser également les cibles de nombre de relevés. Le graphique 1 donne un tracé du nombre optimal de relevés par agglomération à coût de collecte d'ensemble inchangé, pour l'échantillon sélectionné. Le tableau 4 donne le nombre moyen

31. Au final, seules 99 agglos sont retenues, l'une des agglos sélectionnées initialement n'ayant pratiquement aucun point de vente sur son territoire. Le poids w_a de cette agglo a été revu après examen *in situ* et considéré alors comme nul. Toutes les formules dérivées précédemment s'appliquent avec un poids $w_a = 0$ pour cette agglo (et $\pi_a \neq 0$), tout en conservant $\mathcal{N} = 100$.
32. Si la probabilité d'inclusion ainsi obtenue est supérieure à 1, elle est saturée à 1 et le calcul des probabilités d'inclusion des autres agglos est relancé avec un nouveau coefficient de proportionnalité tel que la somme soit unitaire sur les probabilités itérées. Les itérations se poursuivent ainsi jusqu'à ce qu'aucune probabilité d'inclusion calculée ne soit strictement supérieure à 1.
33. Cette proxy est fondée sur les poids en dépense de consommation alimentaire des ménages par type d'agglo (Jaluzot 2014), répartis par agglo, au sein des classes de type d'agglos, en proportion de la variable de population des ménages au dernier RP (population municipale 2010).

TABLE 1 – Nombre d'unités urbaines par strate métropolitaine

Strate		N_h	IPC	optimi-	n_h	sélec.
Taille d'UU	ZISP			sation		auto.
A	11	1	1	1	1	1
B2	08	3	2	3	3	3
	20	6	5	6	6	6
	28	5	5	5	5	5
	36	5	5	5	5	5
	40	5	5	3,62	4	3
	64	7	5	5,25	5	3
B1	08	5	2	2,34	2	1
	20	4	2	1,79	2	1
	28	4	1	1,94	2	1
	36	2	1	1,15	1	1
	40	4	1	1,59	2	1
	64	3	0	1,44	1	0
C	08	28	5	4,18	4	0
	11	14	3	1,65	2	0
	20	29	5	4,22	4	0
	28	32	6	4,77	5	2
	36	18	3	3,03	3	1
	40	29	5	3,69	4	1
	64	38	7	5,32	5	1
D	08	285	10	5,21	5	0
	11	97	3	1,85	2	1
	20	256	4	4,08	4	2
	28	290	5	4,96	5	2
	36	255	6	4,66	5	1
	40	275	5	4,25	4	1
	64	518	15	9,01	9	0
Total		2218	117	100	100	43

Note : ZISP 08=Aquitaine, 11=IdF (y.c. Yonne), 20=NPdC (y.c. Ardennes), 28=Rhône-A. (n.c. Yonne), 36=L-R, 40=Lorraine (n.c. Ardennes), 64=Bretagne ; N_h est le nombre d'unité urbaines de la strate ; IPC est le nombre d'unité urbaines de la strate couvertes par la collecte actuelle de l'IPC ; "optimisation" désigne le nombre d'unités urbaines qu'il conviendrait de retenir dans la strate si le poids de la strate était strictement proportionnel à la variable de dépense de consommation au lieu d'achat ; n_h est le nombre d'unités urbaines (UU) retenu pour la strate en base 2015. La colonne "sélec. auto" indique, parmi les unité urbaines sélectionnées dans l'échantillon base 2015, combien sont sélectionnées avec une probabilité d'inclusion (conditionnelle à leur appartenance à l'échantillon actuel) égale à 1.

TABLE 2 – Poids des regroupements d'unités urbaines utilisés pour la stratification

Strate géographique	Poids (en %)			
	démographique (1)	alimentaire (2)	(1) normalisé sans le rural	(2) normalisé sans le rural
A	16,7	18,1	21,5	19,5
B2	24,6	25,6	31,7	27,6
B1	5,3	7	6,8	7,6
C	13,6	18,6	17,5	20,1
D	17,4	23,4	22,4	25,2
Rural	22,5	7,4		
Total	100	100	100	100

Note : la colonne de poids démographique est issue de la population au sens du RP 2010 et celle de l'alimentaire est issue du traitement de l'enquête BDF en localisation au lieu d'achat (Jaluzot 2014). Les deux dernières colonnes de poids correspondent aux deux premières, respectivement, le poids du rural étant annulé. Ceci revient à faire l'hypothèse que les achats réalisés dans le rural sont effectués dans les unités urbaines, au prorata du poids de ces unités urbaines dans la consommation.

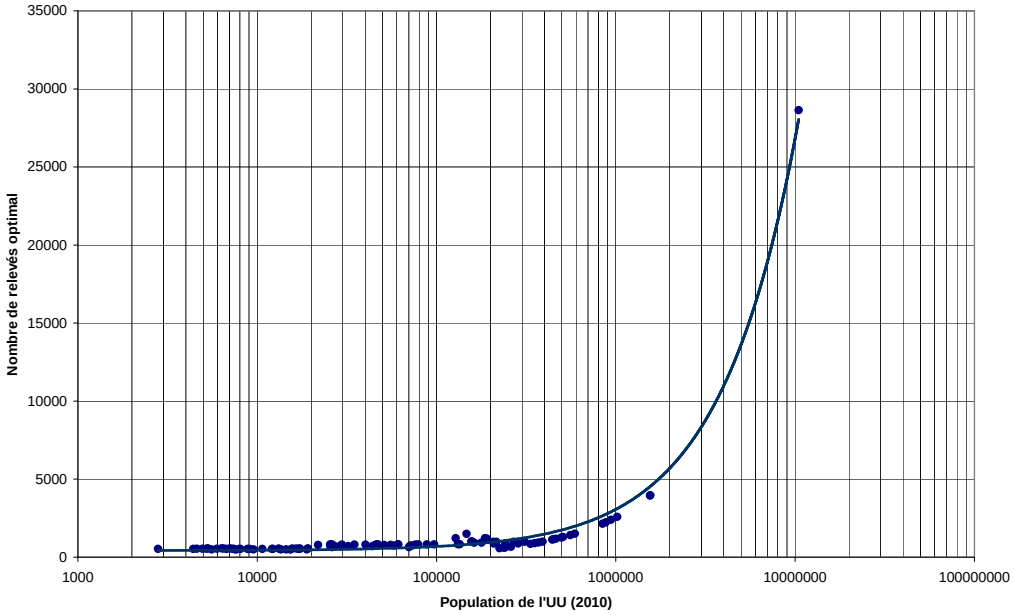
TABLE 3 – Durées moyennes d'observation par secteur de relevé utilisées pour le coût élémentaire de relevé c_v

secteur	durée moyenne hors déplacement	normalisée y.c. dépl. d'un PV à l'autre	Nombre de point de vente par relevé
Bien durables	102s	1.30	0.27
Habillement	60s	0.86	0.19
Alimentaire	49s	0.53	0.09
P. manufacturé	78s	1.16	0.27
Services	25s	1.73	0.61

Note : la colonne 3 comprend une mesure α_i normalisée de temps élémentaires sectoriels issus de la colonne 2. La normalisation est telle que $\sum_{i \in secteur} \alpha_i Nobs_i = 112000$, 112000 relevés étant approximativement le nombre de relevés réalisés au mois de décembre pour l'IPC en 2012 et 2013. Cette mesure tient compte des déplacements d'un point de vente à l'autre lors d'une tournée d'enquêteur. Le temps de déplacement est fixé à 7min13s par point de vente fréquenté. Le nombre de points de vente fréquentés par relevé, selon le secteur, est indiqué en colonne 4. Ces calculs sont issus de l'analyse des temps d'observation élémentaire par relevé (voir Sillard (2012b)).

de relevés par catégorie d'unité urbaine dans l'échantillon actuel et dans l'échantillon-type optimisé en base 2015. Par rapport à la pratique actuelle, le nombre de relevés [34] devrait augmenter sur les petites agglos, passant d'une moyenne de 177 relevés actuellement à une moyenne de 529 relevés pour les unités urbaines de moins de 20 000 habitants (D) ; il devrait également augmenter pour les unités urbaines de 20 000 à 100 000 habitants (C), passant d'une moyenne de 464 à 767 relevés en moyenne ; pour les unités urbaines de 100 000 à 200 000 habitants (B1) et pour les unités urbaines de plus de 200 000 habitants (hors Paris – B2), le nombre d'observations devrait diminuer, comme l'indiquent les chiffres du tableau 4. A noter que pour ces unités urbaines, le nombre optimal de relevés à réaliser est très dépendant de la taille d'unité urbaine, donc l'indication donnée par l'évolution de la moyenne du nombre de relevés entre l'échantillon base 1998 et l'échantillon base 2015 est à relativiser en conséquence [35]. Enfin, sur l'unité urbaine de Paris, le nombre de relevés à réaliser devrait passer de 22 900 à 28 600.

FIGURE 1 – Lien entre nombre optimal de relevés et population de l'unité urbaine



Note : Tracé des couples de points (population de l'UU, nombre optimal de relevés) pour les 100 agglos d'un échantillon type. La courbe est l'approximation linéaire optimale pour ces points. Elle est d'équation $Nb\_releves = 0.0026 \times pop\_mun + 424$. La même relation calculée sur la collecte de l'année 2013 en base 1998 est $Nb\_releves = 0.0022 \times pop\_mun + 456$. *Rem. : Échelle logarithmique pour les abscisses.*

La répartition du nombre de relevés par secteur de produit est également légèrement modifiée par rapport à l'existant. Après optimisation du nombre de relevés par variétés, les principales modifications (cf. tables 4 et 5) portent sur l'alimentaire, où le nombre de relevés devrait passer de 36 000 à 49 000 et sur l'habillement où le nombre de relevés devrait diminuer de 21 000 à 11 000. La masse de relevés consacrés aux autres secteurs est peu modifiée.

34. On rappelle qu'un jour-plein de collecte représente une petite centaine de relevés.
35. Un modèle de dépendance du nombre optimal de relevés à réaliser en fonction de la taille d'unité urbaine est indiqué en note de la figure 1.

TABLE 4 – Nombre de relevés en fonction du type d'agglomération

Type d'agglo	Nb. relevés b. 1998	Nb. relevés b. 2015
A	22862	28633
$B2$	2105 [365 ; 5048]	1389 [570 ; 3966]
$B1$	1485 [1263 ; 1909]	1063 [826 ; 1491]
C	464 [137 ; 1145]	767 [637 ; 820]
D	177 [100 ; 381]	529 [489 ; 562]

Note : les types d'unité urbaines sont – l'unité urbaine de Paris (A) ; – les unités urbaines de plus de 200 000 habitants (B2) hors Paris ; – les unités urbaines de 100 000 à 200 000 habitants (B1) ; – les unités urbaines de 20 000 à 100 000 habitants (C) ; – les unités urbaines de moins de 20 000 habitants (D, hors zones rurales). Les intervalles correspondent aux valeurs extrêmes observées sur les agglos de la catégorie concernée. Calcul réalisé sur les agglos de la base 2015.

TABLE 5 – Nombre de relevés en fonction du type de secteur

Secteur	Base 1998 (total)	Base 1998 (Agglos conservées)	Base 2015
AL	35 568	31 801	49 380
BD	6 544	6 074	5 652
HA	21 033	20 451	11 449
MA	29 939	27 683	29 019
SE	19 683	18 119	21 375
Total	112 767	104 128	116 875

Note : AL=alimentaire, BD=biens durables, HA=habillement, MA=produits manufacturés, SE=services ; "Agglos conservées" désigne les agglos de la base 1998 qui sont conservées en base 2015.

S'agissant des produits frais, ceux-ci sont exclus de l'exercice d'optimisation car leurs relevés ne correspondent pas, pour l'instant, au schéma retenu pour le reste de la collecte IPC. En revanche, le changement de base est l'occasion de faire évoluer la collecte pour la rendre homologue à celle des produits alimentaires [36]. Dans ce contexte, le nombre de relevés de produits frais est déterminé de sorte que l'effort de collecte consenti en base 1998 par l'Insee perdure en base 2015. En effet, les relevés de produits frais sont en particulier utilisés pour la publication de prix moyens par type de produits et l'optimisation du nombre de relevés aux seules fins de publication d'indice de prix ne prendrait pas en compte cette exigence particulière. Il a donc été décidé de maintenir le dimensionnement actuel de l'échantillon. Ceci définit le nombre total de relevés de produits frais à réaliser en nouvelle base (en l'occurrence 37 000 sur la métropole). Ces relevés seront, dans un premier temps, ventilés entre agglos de collecte proportionnellement au nombre de relevés réalisés au titre de l'alimentaire. Une optimisation pourra être réalisée en 2017 sur les mêmes principes que ceux appliqués pour les autres secteurs après l'observation d'une année complète des produits frais conformément à la nouvelle méthodologie.

7.4 Le nombre de relevés par forme de vente et par point de vente

Pour chaque variété de l'IPC, la distribution marginale des ventes par forme de vente [37] est supposée connue de l'expert sectoriel de la division des prix à la consommation en charge du suivi de la variété concernée. Cette connaissance repose sur différentes sources de documentations, notamment les données issues des panélistes. La répartition marginale est ensuite appliquée, par agglo, au nombre de relevés fixé.

Dans le but de ne pas engendrer d'"effets de grappes" dans la mesure de l'évolution des prix, les enquêteurs sont invités à ne pas multiplier les relevés dans un point de vente donné. Ce nombre dépend de la taille du point de vente et a été fixé sur la base de critères statistiques (Division des prix à la consommation 2006). Ce principe est conservé en base 2015. Le nombre de relevés important à réaliser sur de petites agglomérations en base 2015 peut se heurter au faible nombre de points de vente sur site, compte tenu notamment de la contrainte précédente. Selon le cas, il peut donc s'avérer nécessaire d'adapter le calcul statistique à la réalité du terrain. Dans un tel cas, tout se passe comme si le poids de l'agglo en dépense de consommation des ménages avait été initialement sur-évalué. Il convient donc de le revoir à la baisse, en gardant en tête que la probabilité d'inclusion de l'agglo dans l'échantillon est inchangée, c'est-à-dire qu'elle reste dictée par le poids w_a initialement utilisé pour le calcul du plan de sondage. Ainsi, à probabilité d'inclusion de l'agglo a dans l'échantillon inchangée, les résultats issus des formules (11) et (14) sont à recalculer conformément à ces formules après avoir diminué les $w_{a,v}$ pour l'agglo a considérée (et pour des π_a inchangés). Au premier ordre, ceci va se traduire par le maintien du nombre total de relevés à réaliser pour la variété v concernée (variable n_v dans les expressions précédentes) et un "transfert" des relevés en moins dans l'agglo a vers l'ensemble des autres agglos, approximativement en proportion des nombres de relevés initialement calculés pour elles au titre de la variété considérée. Le processus d'adaptation des cibles de nombre de relevés décrit ci-dessus est mis en œuvre quand le terrain l'impose. En pratique, ceci ne concerne qu'un faible nombre de couples (agglos,variété). À noter qu'à l'issue de l'analyse du tissu commercial réalisé par les sites prix, une des 100 agglos initialement sélectionnée s'est avérée dépourvue des principales formes de vente. Il a été décidé d'abandonner cette agglo et de réviser son poids en dépense de consommation des ménages (w_a dans les expressions précédentes) en le fixant à 0.

36. Pour plus de détails sur le fondement de cette évolution et les tests qui ont été réalisés et ont conduit à la définition du nouveau contour méthodologique, se reporter à Corbel (2014).
37. les points de vente de l'IPC sont classés parmi 11 formes de vente (avec leur code entre parenthèses) : Hypermarché (10), Supermarché (20), Maxidiscount (25), Supérette (30), Magasin populaire (40), Grand magasin (50), Grande surface spécialisée (60), Magasin traditionnel (70), Marché (80), Services (90), autres (99).

8 De la base 1998 à la base 2015

En pratique, on doit faire évoluer un échantillon de collecte de la base 1998 à celui de la base 2015 tout en :

1. continuant de produire l'indice, en base 1998, jusque fin 2015, puis en base 2015 à partir de janvier 2016 ;
2. permettant un calcul en base 2015 sur l'année 2015 destiné à fixer le coefficient de chaînage de sorte que la moyenne de l'indice chaîné-base 2015 soit égale à 100 sur l'année 2015.

Sur les 120 000 produits suivis dans l'IPC métropolitain en 2014 (base 1998), près de la moitié auront été renouvelés lors du changement de base, pour l'essentiel, *du fait* du changement de base.

Il est évident qu'une telle opération ne peut pas s'effectuer immédiatement : les enquêteurs ne sont pas instantanément mobiles et la modification de zones d'enquête nécessite une coordination soigneuse. Nous avons opté pour un processus par étapes s'appuyant sur le processus annuel d'évolution d'échantillon opéré annuellement lors des *opérations de changement d'année*. En effet, l'IPC est un indice chaîné annuellement dont l'échantillon évolue d'une année sur l'autre pour rendre compte des évolutions de consommation des ménages. Mais les évolutions annuelles classiques restent limitées à environ 5% de l'échantillon. Dans le cas du changement de base, les modifications portent sur près de 50% de l'échantillon. Nous avons décidé de procéder à ces évolutions en trois années en s'appuyant sur le processus mis en œuvre lors des opérations de changement d'année 2015 (vers 2016), 2016 (vers 2017) et 2017 (vers 2018).

La collecte IPC est administrée par 7 sites prix métropolitains qui ont une vue détaillée de l'activité de chaque enquêteur. Les sites prix ont tout d'abord validé la faisabilité d'atteinte des cibles de nombre de relevés par agglomérations de la base 2015, le cas échéant en les modifiant lorsque le tissu commercial ne permettait pas d'atteindre le nombre de relevés établi par optimisation, tout en appliquant les règles de nombre de relevés maximal fixé par forme de vente pour éviter les effets de grappes (voir §7.4). Puis, un programme triennal d'atteinte des cibles de la base 2015 a été établi en fonction des moyens disponibles, l'Insee ayant décidé de consacrer un surcroît de moyens de collecte à l'IPC de 7% sur l'année 2015. En effet, cette année-là, une double collecte est nécessaire puisque d'une part, il convient de collecter l'information nécessaire au calcul de l'IPC en base 1998 et d'autre part, il convient de collecter l'information utile au calcul de l'indice en base 2015. Le volume de surcroît de collecte a été déterminé une fois connu l'échantillon d'agglomérations en base 2015 issu de l'échantillonnage présenté au paragraphe 7.

Le programme triennal s'appuie sur deux *inputs* : d'une part un nombre de relevés à atteindre pour l'échantillon de l'année concernée en base 2015 par agglo ; d'autre part, un nombre de relevés par variété métropolitaine fixé par les équipes sectorielles de la division des prix à la consommation, au vu des résultats de l'optimisation du nombre de relevés par variété-agglomération effectuée comme indiqué au paragraphe 7.3.

Chaque année du programme triennal (2015, 2016, 2017), on fixe le nombre de produits à créer par "ordre de recherche" de sorte que le nombre d'ordres de recherche soit minimal sous la double contrainte que (1) le nombre de relevés par agglo soit égal au nombre fixé dans le programme triennal et (2) que le nombre de relevés par variété soit égal au nombre fixé par les équipes sectorielles de la division des prix à la consommation. La mise en œuvre de cette optimisation permettant de déterminer le nombre d'ordres de recherche à réaliser par var-agglo est détaillée à l'annexe G.

Les résultats de l'optimisation fixant le nombre de relevés par variété-agglomération sont des nombres décimaux. Il convient donc d'arrondir ces nombres à des entiers. Cependant, arrondir le nombre de relevés au plus proche

entier n'est pas satisfaisant, car il peut entrainer un biais par rapport aux contraintes fixées dans le calcul [38]. La méthode utilisée pour passer d'un nombre décimal de relevés à un nombre entier en respectant, en moyenne, le nombre de relevés fixés sur les marges consiste à appliquer une méthode probabiliste de détermination du nombre de relevés [39]. En première étape, cela consiste à utiliser une loi de Bernouilli de la manière suivante. Soit $x^{\star}_{av}$ le nombre de relevé décimal optimal pour l'agglo a et la variété v. Le nombre de relevés retenu est : $n^{\star}_{av} = \mathbf{E}(x^{\star}_{av}) + y$ où $\mathbf{E}$ désigne la partie entière et y est une réalisation d'une variable tirée selon une loi de Bernouilli $\mathscr{B}(x^{\star}_{av} - \mathbf{E}(x^{\star}_{av}))$. Avec cette méthode et en application de la loi des grands nombres, les marges fixées sont respectées en moyenne.

L'annexe G précise le détail des règles d'arrondis réalisées, notamment en distinguant selon le type de variété concerné.

38. Imaginons, par exemple, que pour une variété v donnée, tous les nombres décimaux s'établissent à $x_{av} = 0,3$ relevés pour l'ensemble des varagglos. En moyenne il faudrait donc 30 relevés pour cette variété ($0,3 \times 10\ agglos = 30$). Mais arrondir au plus proche entier conduirait à ne retenir aucun relevé par agglo, donc *in fine*, 0 relevé pour l'ensemble de la variété, c'est-à-dire un nombre très éloigné de la contrainte fixée.

39. C'est d'ailleurs ce genre de méthode qui est utilisée par l'algorithme de calcul le tableau d'équilibrage par forme de vente dans les applications IPC (Viglino 2004).

Références

Ardilly, P. (2006). *Les techniques de sondage*, 2 edn, Technip.

Ardilly, P. (2014). Échanges entre la division des prix et M. Pascal Ardilly à propos de la sélection des agglomérations IPC dans la Base 2015, *Communication personnelle, Note interne N° 295/DG75-F320/*, Insee, division des prix à la consommation.

Ardilly, P. & Guglielmetti, F. (1992). Optimisation de l'échantillon pour le calcul de l'indice de prix à la consommation, *Insee méthodes* **29-30-31** : 71 – 123.

Ardilly, P. & Guglielmetti, F. (1993). La précision de l'indice des prix : mesure et optimisation, *Économie et Statistique* **267** : 13 – 29.

Balcone, T. (2014). La précision de l'indice des prix à la consommation, *Note interne N° 1460/DG75-F320/TB*, Insee, division des prix à la consommation.

Corbel, P. (2014). Note de méthodologie pour le calcul de l'indice des produits frais en nouvelle base, *Note interne N° 193/DG75-F320/*, Insee, division des prix à la consommation.

Division des prix à la consommation (2006). Guide de l'enquêteur des prix à la consommation, *Manuel de collecte*, INSEE.

Faivre, S. (2012). Calcul de l'échantillon d'agglomérations optimal pour l'indice des prix suite au passage aux données de caisses, *7ème colloque francophone sur les sondages, ENSAI - Rennes, 5 – 7 novembre 2012*.

Jaluzot, L. (2014). Étude de la géographie de la consommation des ménages, *Note interne N° 172/DG75-F320/LJ*, Insee, division des prix à la consommation.

Oehlert, G. W. (1992). A Note on the Delta Method, *The American Statistician* **46**(1) : 27 – 29.

Petit, G. (2014). Précision de l'indice des prix à la consommation, *Note interne N° 815/DG75-F320/*, Insee, division des prix à la consommation.

Sillard, P. (2012a). Proposition de calcul de quotité des enquêteurs-prix métropolitains, *Note interne N° 887/DG75-F320/PS*, Insee, division des prix à la consommation.

Sillard, P. (2012b). Temps d'observations modélisés pour les enquêteurs prix, *Note interne N° 1523/DG75-F320/PS*, Insee, division des prix à la consommation.

Tillé, Y. (2001). *Théorie des sondages*, Dunod.

Viglino, L. (2004). Spécifications concernant l'algorithme qui détermine les objectifs par case des tableaux d'équilibrage , *Note interne N° 130/F320*, Insee, division des prix à la consommation.

A Les formules d'agrégation utilisées dans l'IPC en base 1998

La table 6 précise dans quelles tables des applications IPC se situent les variables mentionnées dans cette partie.

A.1 L'agrégation des indices de variétés et au-delà

L'agrégation des indices de variétés I_v s'opère selon une formule de Laspeyres :

$$I = \sum_v POND_v \times I_v$$

où $POND_v$ est un poids en valeur. Ce poids est obtenu par la composition d'un poids en valeur du [poste,territoire], issu de la comptabilité nationale et noté $POND_P$, et d'un poids de la variété dans le poste $POSTE$. Par hypothèse,

$$\sum_{v \in P} POIDS_v = 100$$

et

$$POND_v = \frac{POIDS_v}{100} \times POND_P$$

A.2 Le calcul de l'indice de variété I_v

Par hypothèses, étant donnée une série d'indice $I_{v,a}$ de varagglo, l'indice de variété se calcule comme suit :

$$I_v = \sum_a POND_{v,a} \times I_{v,a}$$

Le détail des calculs de la variable $POND_{v,a}$ est donné aux paragraphes suivants.

A.2.1 Le calcul de $POND_{v,TA}$

On définit d'abord le poids de la variété dans le type d'agglo ($TA \in \{A, B, C, D\}$) :

$$POND_{v,TA} = \frac{POIDS_{v,TA}}{100}.\frac{POIDS_v}{100}.POND_{P,terri}$$

- les variables $POIDS_{v,TA}$ et $POIDS_v$ sont fixées par le sectoriel de la division des prix en charge du suivi de la variété v, lors de la définition de la variété (opérations de changement d'année). Ainsi défini, $\sum_{TA \in \{A,B,C,D\}} POIDS_{v,TA} = 100$ caractérise la répartition de la dépense de la consommation de la variété v sur le territoire considéré selon les 4 types d'agglos $\{A, B, C, D\}$.
- $POND_{P,terri}$ est calculé à partir des éléments de la Comptabilité nationale.

Avec cette formule, on répartit le poids du [poste×terri] sur chaque couple (v, TA). $POND_{v,TA}$ est donc un poids en dépense, de sorte que :

$$\sum_{\substack{v \in POSTE \\ TA \in \{A,B,C,D\}}} POND_{v,TA} = POND_{P,terri}$$

A.2.2 Calcul de $POND_{v,a}$ pour le type d'agglo=A

$$POND_{v,a=PARIS} = POND_{v,TA=A}$$

A.2.3 Calcul de $POND_{v,a}$ pour le type d'agglo=B

On répartit $POND_{v,TA}$ sur l'ensemble des agglos $a \in \{TA = B\}$ selon un coefficient de répartition :

$$POND_{v,a} = R_{v,a} \times POND_{v,TA=B}$$

On définit d'abord le nombre de relevé réalisés dans l'agglo a :

$$NBENQ_a = \sum_v NBENQ_{v,a}$$

et

$$R_{v,a} = \frac{POIDS_a \times NBENQ_{v,a}/NBENQ_a}{\sum_{a \in \{TA=B\}} POIDS_a \times NBENQ_{v,a}/NBENQ_a}$$

A.3 Calcul de $POND_{v,a}$ pour le type d'agglo={C,D}

La pondération $POND_{v,TA}$ sur l'ensemble des agglos de type $TA \in \{C, D\}$ est répartie entre agglos selon une formule

$$POND_{v,a} = R_{v,a} \times POND_{v,TA}$$

où

$$R_{v,a} = \frac{NBENQ_{v,a}}{NBENQ_{v,TA}}$$

C'est le poids relatif de l'agglo a dans sa taille d'agglo, en nombre de relevés effectués au titre de la variété considérée.

TABLE 6 – Liste des variables et tables des applications IPC correspondantes

Table IPCNAT	variable	colonne
AGGLO	$POIDS_a$	POIDS
ANVA	$POND_{v,a}$	POND
"	$NBENQU_{v,a}$	NBENQ
VARTA	$POIDS_{v,TA}$	POIDS
"	$NBENQ_{v,TA}$	NBENQ
VARIETE	$POIDS_v$	POIDS
"	$POND_v$	PONDTOT

B Optimisation du nombre de relevés

On cherche à résoudre le programme d'optimisation suivant :

$$\boldsymbol{\nu}^\star = \underset{\boldsymbol{\nu}}{\operatorname{argmin}} \left\{ \sum_i \frac{a_i^2}{\nu_i} \middle| \sum_i \alpha_i^2 \nu_i = K \right\}$$

où $\boldsymbol{\nu}$ est un vecteur d'inconnues de dimension M ($\boldsymbol{\nu} = (\nu_1, \ldots, \nu_M)$) et les $(a_i)_{1 \leqslant i \leqslant M}$, $(\alpha_i)_{1 \leqslant i \leqslant M}$ et K sont des paramètres du programme. Les a_i sont positifs ou nuls ; les α_i et K sont strictement positifs.

Le lagrangien du programme s'écrit (λ est le multiplicateur de Lagrange) :

$$\mathscr{L}(\boldsymbol{\nu}, \lambda) = \sum_i \frac{a_i^2}{\nu_i} + \lambda(\sum_i \alpha_i^2 \nu_i - K)$$

Puis,

$$\frac{\partial \mathscr{L}}{\partial \nu_i} = -\frac{a_i^2}{\nu_i^2} + \lambda \alpha_i^2$$

Les conditions de premier ordre $\frac{\partial \mathscr{L}}{\partial \nu_i} = 0$ conduisent à :

$$\lambda = \frac{a_i^2}{\alpha_i^2 \nu_i^2}$$

Il en découle que :

$$\nu_i = \frac{a_i}{\alpha_i \sqrt{\lambda}}$$

Or $\sum_i \alpha_i^2 \nu_i = K$ donc,

$$\sqrt{\lambda} = \frac{1}{K} \sum_i \alpha_i a_i$$

Finalement, pour tout $i \in \{1, \ldots, M\}$,

$$\nu_i^\star = \frac{K}{\alpha_i^2} \times \frac{\alpha_i a_i}{\sum_i \alpha_i a_i}$$

C Variance des micro-indices et biais de formule à distance finie

C.1 Variétés homogènes

On considère un ensemble de bien de taille N. Chaque bien est indicé par $i \in \{1, \ldots, N\}$. Le prix des biens est p_i^t à l'instant t et p_i^0 à l'instant 0. On cherche à estimer l'indice

$$I = \frac{\frac{1}{N} \sum_{i=1}^N p_i^t}{\frac{1}{N} \sum_{i=1}^N p_i^0}$$

en sélectionnant par sondage aléatoire simple avec remise n produits. On note $\mathscr{S}$ cet échantillon et on suppose que le taux de sondage n/N est très petit devant l'unité, de sorte qu'on peut le négliger. On estime I par $\hat{I}$ défini par :

$$\hat{I} = \frac{\frac{1}{n} \sum_{i \in \mathscr{S}} p_i^t}{\frac{1}{n} \sum_{i \in \mathscr{S}} p_i^0}$$

La variance de $\hat{I}$ provient de la variance des deux variables $\hat{\bar{P}}^t = \frac{1}{n} \sum_{i \in \mathscr{S}} p_i^t$ et $\hat{\bar{P}}^0 = \frac{1}{n} \sum_{i \in \mathscr{S}} p_i^0$ en tant qu'estimateurs de $\bar{P}^t = \frac{1}{N} \sum_{i \in \{1\ldots,N\}} p_i^t$ et $\bar{P}^0 = \frac{1}{N} \sum_{i \in \{1\ldots,N\}} p_i^0$. On sait calculer $\mathrm{var}(\hat{\bar{P}}^t)$ et trouver un estimateur $\widehat{\mathrm{var}(\hat{\bar{P}}^t)}$ de cette variance à l'aide des prix observés sur l'échantillon $\mathscr{S}$. De même s'agissant de $\hat{\bar{P}}^0$. Il reste donc à calculer $\mathrm{var}(\hat{I})$ étant données les variance de $\hat{\bar{P}}^t$ et de $\hat{\bar{P}}^0$.

Classiquement, on applique le principe de la Δ-méthode : si $\bar{X}_n$, $\bar{Y}_n$ et $\bar{Z}_n$ sont des variables aléatoires telles que $\bar{Z}_n = F(\bar{X}_n, \bar{Y}_n)$ où F est une fonction deux fois dérivable et $(\bar{X}_n, \bar{Y}_n)$ est un vecteur de variables

aléatoires – moyennes arithmétiques de variables iid (de moyenne $(\mathbb{E}X, \mathbb{E}Y)$) sur des échantillons de taille n, alors, en développant F au voisinage de $(\mathbb{E}X, \mathbb{E}Y)$ on a [40] :

$$\begin{aligned} \bar{Z}_n &= F(\mathbb{E}X, \mathbb{E}Y) + \partial_1 F(\mathbb{E}X, \mathbb{E}Y).(\bar{X}_n - \mathbb{E}X) + \\ &\quad \partial_2 F(\mathbb{E}X, \mathbb{E}Y).(\bar{Y}_n - \mathbb{E}Y) + \bar{\varepsilon}_n \end{aligned} \tag{22}$$

On montre [41] que $\mathrm{var}(\bar{\varepsilon}_n) = o(1/n)$. Il en découle que [42] :

$$\begin{aligned} \mathrm{var}(\bar{Z}_n) &= \left[\partial_1 F(\mathbb{E}X, \mathbb{E}Y)\right]^2 .\mathrm{var}(\bar{X}_n) + \left[\partial_2 F(\mathbb{E}X, \mathbb{E}Y)\right]^2 .\mathrm{var}(\bar{Y}_n) \\ &\quad +2\left[\partial_1 F(\mathbb{E}X, \mathbb{E}Y)\right].\left[\partial_2 F(\mathbb{E}X, \mathbb{E}Y)\right]\mathrm{cov}(\bar{X}_n, \bar{Y}_n) + o(1/n) \end{aligned} \tag{23}$$

Dans le cas des variétés homogènes, $F(X,Y) = \frac{X}{Y}$ donc $\partial_1 F(X,Y) = \frac{1}{Y}$ et $\partial_2 F(X,Y) = -\frac{X}{Y^2}$. Avec les notations précédentes, nous avons :

$$\mathrm{var}(\hat{I}) = \left(\frac{1}{\bar{P}^0}\right)^2 \mathrm{var}(\hat{\bar{P}}^t) + \left(\frac{\bar{P}^t}{(\bar{P}^0)^2}\right)^2 \mathrm{var}(\hat{\bar{P}}^0) - 2\frac{1}{\bar{P}^0} \times \frac{\bar{P}^t}{(\bar{P}^0)^2}\mathrm{cov}(\hat{\bar{P}}^t, \hat{\bar{P}}^0)$$

On remarque que :

$$\mathrm{var}(\hat{I}) = \left(\frac{1}{\bar{P}^0}\right)^2 \mathrm{var}\left(\hat{\bar{P}}^t - \frac{\bar{P}^t}{\bar{P}^0}\hat{\bar{P}}^0\right)$$

Moyennant quoi, il ne reste plus qu'à construire un estimateur de l'expression précédente en remarquant que :

$$\hat{\bar{P}}^t - \frac{\bar{P}^t}{\bar{P}^0}\hat{\bar{P}}^0 = \frac{1}{n}\sum_{i\in\mathscr{S}}\left(p_i^t - \frac{\bar{P}^t}{\bar{P}^0}p_i^0\right)$$

Un estimateur de la variance de cette moyenne empirique est classiquement :

$$\widehat{\mathrm{var}\left(\hat{\bar{P}}^t - \frac{\bar{P}^t}{\bar{P}^0}\hat{\bar{P}}^0\right)} = \frac{1}{n}\frac{1}{n-1}\sum_{i\in\mathscr{S}}\left[p_i^t - \frac{\hat{\bar{P}}^t}{\hat{\bar{P}}^0}p_i^0 - \left(\hat{\bar{P}}^t - \frac{\hat{\bar{P}}^t}{\hat{\bar{P}}^0}\hat{\bar{P}}^0\right)\right]^2$$

Soit, après simplifications (voir Eq. 16) :

$$\widehat{\mathrm{var}(\hat{I})} = \left(\frac{1}{\hat{\bar{P}}^0}\right)^2 \frac{1}{n(n-1)}\sum_{i\in\mathscr{S}}\left(p_i^t - \hat{I}p_i^0\right)^2$$

Il est possible d'évaluer le biais de formule de l'indice, c'est-à-dire l'écart entre $\mathbb{E}(\hat{I})$ et I. Ce biais dépend directement de la taille n de l'échantillon $\mathscr{S}$. En effet, si on développe $\bar{Z}_n$ à l'ordre 2, nous avons :

$$\begin{aligned} \bar{Z}_n &= F(\mathbb{E}X, \mathbb{E}Y) + \partial_1 F(\mathbb{E}X, \mathbb{E}Y).(\bar{X}_n - \mathbb{E}X) + \partial_2 F(\mathbb{E}X, \mathbb{E}Y).(\bar{Y}_n - \mathbb{E}Y) + \\ &\quad +\tfrac{1}{2}\left[\partial_{11} F(\mathbb{E}X, \mathbb{E}Y).(\bar{X}_n - \mathbb{E}X)^2 + \partial_{22} F(\mathbb{E}X, \mathbb{E}Y).(\bar{Y}_n - \mathbb{E}Y)^2 + \right. \\ &\quad \left. + 2\partial_{12} F(\mathbb{E}X, \mathbb{E}Y).(\bar{X}_n - \mathbb{E}X)(\bar{Y}_n - \mathbb{E}Y)\right] + \bar{\xi}_n \end{aligned} \tag{24}$$

On montre[41] que $\mathbb{E}(\bar{\xi}_n) = o(1/n)$. Moyennant quoi, pour les variétés homogènes [43], le biais relatif s'écrit [44] :

$$\frac{\mathbb{E}(\hat{I}) - I}{I} = \left(\frac{\sigma}{\bar{P}^0}\right)^2 \times \frac{I-\rho}{I} \times \frac{1}{n} + o(1/n) \tag{25}$$

où σ^2 est la variance caractérisant la dispersion des p_i^0 (supposés iid) autour de $\bar{P}^0$ et ρ est la corrélation des variables de prix $(p_i^0, p_i^t)_{i\in\{1,\ldots,n\}}$.

40. $\partial_i F$ désigne la dérivée partielle de F par rapport à sa $i^{\text{ème}}$ composante.
41. Voir par exemple (Oehlert 1992).
42. Compte tenu des hypothèses, $\mathrm{var}(\bar{X}_n) = \sigma_X^2/n$ où σ_X^2 est la variance de X ; une propriété similaire est vérifiée par Y.
43. $\partial_{11}F(X,Y) = 0$, $\partial_{12}F(X,Y) = -1/Y^2$ et $\partial_{22}F(X,Y) = 2\,X/Y^3$.
44. On suppose que $\mathrm{cov}(\bar{X}_n, \bar{Y}_n) = 0$.

C.2 Variétés hétérogènes

Avec les mêmes notations qu'au paragraphe C.1, on cherche cette fois à estimer l'indice

$$I = \frac{\left(\prod_{i=1}^{N} p_i^t\right)^{1/N}}{\left(\prod_{i=1}^{N} p_i^0\right)^{1/N}}$$

On estime I par $\hat{I}$ défini par :

$$\hat{I} = \frac{\left(\prod_{i\in\mathscr{S}} p_i^t\right)^{1/n}}{\left(\prod_{i\in\mathscr{S}} p_i^0\right)^{1/n}} = \exp\left\{\frac{1}{n}\sum_{i\in\mathscr{S}} \ln p_i^t - \frac{1}{n}\sum_{i\in\mathscr{S}} \ln p_i^0\right\}$$

Pareillement, $\mathrm{var}(\hat{I})$ provient de la variance des composantes aléatoires X et Y de la fonction $F(X,Y) = \exp(X-Y)$, où $X = \frac{1}{n}\sum_{i\in\mathscr{S}} \ln p_i^t$ et $Y = \frac{1}{n}\sum_{i\in\mathscr{S}} \ln p_i^0$ sont des estimateurs de $\overline{LP}^t = \frac{1}{N}\sum_{i=1}^{N} \ln p_i^t$ et $\overline{LP}^0 = \frac{1}{N}\sum_{i=1}^{N} \ln p_i^0$. A l'aide d'une Δ-méthode, en utilisant le développement (23) nous avons dans le cas présent ($\partial_1 F(X,Y) = \exp(X-Y)$ et $\partial_2 F(X,Y) = -\exp(X-Y)$) :

$$\begin{aligned}\mathrm{var}(\hat{I}) &= I^2\mathrm{var}(\hat{\overline{LP}}^t) + I^2\mathrm{var}(\hat{\overline{LP}}^0) - 2I^2\mathrm{cov}(\hat{\overline{LP}}^t, \hat{\overline{LP}}^0) \\ &= I^2\mathrm{var}(\hat{\overline{LP}}^t - \hat{\overline{LP}}^0)\end{aligned}$$

Or

$$\hat{\overline{LP}}^t - \hat{\overline{LP}}^0 = \frac{1}{n}\sum_{i\in\mathscr{S}} \left(\ln p_i^t - \ln p_i^0\right)$$

donc

$$\widehat{\mathrm{var}\left(\hat{\overline{LP}}^t - \hat{\overline{LP}}^0\right)} = \frac{1}{n}.\frac{1}{n-1}\sum_{i\in\mathscr{S}}\left[\ln\left(\frac{p_i^t}{p_i^0}\right) - (\hat{\overline{LP}}^t - \hat{\overline{LP}}^0)\right]^2$$

Or $\hat{\overline{LP}}^t - \hat{\overline{LP}}^0 = \ln\hat{I}$. Finalement (voir aussi Eq. (17)),

$$\widehat{\mathrm{var}(\hat{I})} = \frac{\hat{I}^2}{n(n-1)}\sum_{i\in\mathscr{S}}\left[\ln\left(\frac{p_i^t}{p_i^0}\right) - \ln\hat{I}\right]^2$$

De même que pour les variétés homogènes, il est possible d'évaluer le biais de formule d'indice pour les variétés hétérogènes, c'est-à-dire l'écart entre $\mathbb{E}(\hat{I})$ et I. Ce biais dépend directement de la taille n de l'échantillon $\mathscr{S}$. On repart du développement (24). Pour les variétés hétérogènes [45], le biais relatif s'écrit [46] :

$$\frac{\mathbb{E}(\hat{I}) - I}{I} = (1-\tilde{\rho})\tilde{\sigma}^2\frac{1}{n} + o(1/n) \tag{26}$$

où $\tilde{\sigma}^2$ est la variance caractérisant la dispersion des $\ln p_i$ (supposés iid) autour de $\overline{LP}$ et $\tilde{\rho}$ est la corrélation des variables de prix $(\ln p_i^0, \ln p_i^t)_{i\in\{1,\ldots,n\}}$.

45. $\partial_{11}F(X,Y) = \exp(X-Y)$, $\partial_{12}F(X,Y) = -\exp(X-Y)$ et $\partial_{22}F(X,Y) = \exp(X-Y)$.
46. On suppose que $\mathrm{cov}(\bar{X}_n, \bar{Y}_n) = 0$ et que $\tilde{\sigma}^2$ est la même en 0 et en t.

C.3 La pratique du biais de formule

Le terme de biais indiqué aux formules (25) et (26) est généralement assez petit car ρ est assez proche de 1 (de même que I pour les variétés homogènes) et le nombre d'observations par agrégat est relativement grand. Pour les variétés de l'IPC, la valeur médiane [47] du coefficient $(\sigma/\bar{P}^0)^2 \times (I-\rho)/I$ pour les variétés homogènes est de 0,59% ; celle du coefficient $(1-\tilde{\rho})\tilde{\sigma}^2$ pour les variétés hétérogènes est de 0,62%. La pratique actuelle de l'IPC est d'imposer un minimum de 4 produits par agrégat élémentaire var-agglo pour les variétés hétérogènes, à la fois pour limiter l'ampleur du biais de formule mais aussi pour assurer un niveau d'information statistique minimale par agrégat élémentaire. Pour les variétés homogènes, ce minimum est fixé à 2 produits.

Ce faisant, le biais d'indice est limité à 0,3% pour les indices de var-agglo homogènes et 0,15% pour les indices de var-agglos hétérogènes. Le biais, au niveau var-agglo, contribue à un biais d'ensemble en étant affecté du poids de la var-agglo en dépense de consommation des ménages. Donc une agglo de faible poids peut comporter de petits échantillons sans que l'indice calcul d'ensemble ne soit affecté d'un biais prononcé. Et en pratique, c'est ce schéma qui prévaut : les agglos à faible poids sont davantage sujettes à petits échantillons que les agglos pesant fortement dans l'indice d'ensemble.

Une évaluation du biais affectant l'indice d'ensemble en raison du biais de formule à distance finie a été réalisée sur l'année 2013 et pour le champ de la collecte terrain en métropole. Elle aboutit à un biais positif de 0,05%. Ce biais est inférieur à la précision actuelle de l'indice (0,1%). Toutefois, il conviendra d'y porter un certaine attention à mesure que cette précision s'améliore avec l'optimisation de l'échantillon. En effet, la borne inférieure de la précision d'indice atteignable à moyens constants est de 0,02%, soit une valeur moins élevée que le biais d'indice actuel.

C.4 Conséquence sur la précision de l'IPC de la saturation du nombres de relevés par varagglos

Jusqu'à présent, on a fait l'hypothèse que les probabilités d'inclusion π_a d'une agglo dans l'échantillon observé étaient les mêmes pour toutes les variétés. Il découle du principe exposé au paragraphe C.3 et du mécanisme de détermination du nombre de relevés par varagglo (voir §G.3) que ce n'est pas le cas : on préférera, pour une variété donnée, concentrer les relevés sur certaines agglos afin d'atteindre un nombre de relevés tel que l'indice varagglo soit peu affecté par le biais de formule à distance finie (voir §C) plutôt que de multiplier les petits échantillons varagglos, par construction davantage sujets au biais de formule à distance finie.

Un moyen de rendre compte de cette concentration sur le calcul de variance de l'indice est de considérer que le terme de la somme figurant dans l'indice (3) est multiplié par une nouvelle variable aléatoire $\mathbf{1}_v(a)$ qui suit une loi de Bernouilli. En première approximation, on peut considérer que la relation (3) s'écrit, en l'occurrence [48] :

$$\hat{I}_v^{\mathscr{A}} = \sum_{a \in A} \omega_a \hat{I}_{a,v} \times \mathbf{1}_{\mathscr{A}}(a) \times \mathbf{1}_v(a)$$

où $\mathbf{1}_v(a)$ vaut 1 si, pour l'agglo a de l'échantillon $\mathscr{A}$, la varagglo est effectivement observée (i.e. le nombre de produits suivis est non nul) et 0 sinon.

47. Calcul réalisé sur l'échantillon IPC métropolitain de 2013.
48. Comme indiqué au début du paragraphe 3, les notations de ce paragraphe ne font pas apparaître explicitement que l'ensemble des expressions se réfère à une variété donnée, l'indice de variété étant sciemment omis dans les expressions. À l'inverse, dans cette annexe, on ré-introduit l'indice identifiant de la variété pour plus de clarté, car les modifications apportées aux formules liées au phénomène de saturation décrit plus haut dépendent de la variété.

Tout se passe donc comme si, pour la variété v, l'échantillon $\mathscr{A}$ était remplacé par un échantillon $\mathscr{A}_v$ et la formule (4) remplacée par :

$$\hat{I}_v^{\mathscr{A}_v} = \sum_{a \in \mathscr{A}_v} \frac{w_a}{\pi_{a,v}} \hat{I}_{a,v}$$

où $\pi_{a,v} = \pi_a \times \xi_v(a)$, π_a étant la probabilité d'inclusion de a dans $\mathscr{A}$ et $\xi_v(a)$ la probabilité que la varagglo (a, v) fasse l'objet d'un nombre de relevés non nul.

Pour une variété v donnée, $\xi_v(a)$ est soit connue en fonction de l'algorithme utilisé pour décider si le nombre de relevés dans une agglo est égal à 0 (voir paragraphes G.3.3 et G.3.4) ou pas, soit il peut être estimé empiriquement *ex-post*, sous les hypothèses que $\mathbf{1}_{\mathscr{A}}(a)$ et $\mathbf{1}_v(a)$ sont indépendantes, de même que $\mathbf{1}_v(a)$ et $\mathbf{1}_v(a')$ pour $a \neq a'$ et que $\xi_v(a)$ ne dépend[49] pas de a. Un estimateur de ξ_v est :

$$\hat{\xi}_v = \frac{[\text{nombre d'agglos de } \mathscr{A} \text{ dont le nombre de relevés est non nul pour la variété } v]}{[\text{nombre d'agglos de } \mathscr{A}]}$$

Moyennant quoi, la variance du second degré (formule 5) est augmentée d'un coefficient multiplicateur dépendant de la variété et égal à $1 \Big/ \hat{\xi}_v$.

D Calcul de la probabilité d'inclusion optimale pour conserver un maximum d'agglo dans l'échantillon nouveau

On repart de la formule (21) :

$$\pi_a^{\mathscr{X}} = \pi_a^{\mathscr{I}} \times \Pr(a \in \mathscr{X} | a \in \mathscr{I}) + (1 - \pi_a^{\mathscr{I}}) \times \Pr(a \in \mathscr{X} | a \notin \mathscr{I})$$

et on pose :

$$\begin{cases} x &= \Pr(a \in \mathscr{X} | a \notin \mathscr{I}) \\ y &= \Pr(a \in \mathscr{X} | a \in \mathscr{I}) \end{cases}$$

On cherche (x, y) se sorte que y soit maximal (dans l'intervalle $[0, 1]$) et

$$\pi_a^{\mathscr{X}} = \pi_a^{\mathscr{I}}.y + (1 - \pi_a^{\mathscr{I}}).x$$

tandis que x est également dans l'intervalle $[0, 1]$. Tout ceci revient à maximiser la fonction ($y \equiv \varphi(x)$ dans l'expression précédente) :

$$\varphi(x) = \frac{\pi_a^{\mathscr{X}}}{\pi_a^{\mathscr{I}}} - \frac{1 - \pi_a^{\mathscr{I}}}{\pi_a^{\mathscr{I}}} x$$

φ est décroissante. Pour $x \in [0, 1]$, cette fonction est maximale au point $x = 0$. Or $\varphi(0) = \pi_a^{\mathscr{X}} \big/ \pi_a^{\mathscr{I}}$. Mais $y \in [0, 1]$. Par conséquent, deux situations se présentent :

— Si $\pi_a^{\mathscr{X}} \leqslant \pi_a^{\mathscr{I}}$ alors la solution

$$(x, y) = \left(0, \frac{\pi_a^{\mathscr{X}}}{\pi_a^{\mathscr{I}}}\right)$$

est optimale.

— Si $\pi_a^{\mathscr{X}} > \pi_a^{\mathscr{I}}$, alors le maximum de $\varphi(x)$ conditionnellement au fait que $\varphi(x) \leqslant 1$ est atteint au point $\varphi^{-1}(1)$ en lequel φ vaut 1. L'optimum est alors :

$$(x, y) = \left(\frac{\pi_a^{\mathscr{X}} - \pi_a^{\mathscr{I}}}{1 - \pi_a^{\mathscr{I}}}, 1\right)$$

49. Cette dernière hypothèse est fausse puisque naturellement, la probabilité que le nombre de relevés soit nulle est liée à la taille de l'agglomération. Cette approche ne constitue donc qu'une approximation.

E Le cas des DOMs

Chaque territoire DOM est caractérisé, dans l'IPC base 1998, par un unique agrégat géographique. En d'autres termes, partant de prix relevés dans des points de vente du territoire, on ne procède pas au calcul d'agrégat par unité urbaine, mais au calcul d'un agrégat d'ensemble pour le territoire. Il en découle que le nombre de ces agrégats élémentaires, correspondant dans le cas de la métropole aux unités urbaines, est égal à 1 pour chaque territoire DOM. En conséquence, il n'y a pas, dans le cas des DOM, de sondage géographique : par hypothèse, les points de vente retenus pour les observations de l'indice des prix sont représentatifs de l'ensemble des points de vente du territoire et doivent, à ce titre, être répartis sur l'ensemble du territoire.
Ce choix d'un unique agrégat géographique est lié à la nature des biais de formule qui surviennent lorsque peu de produits sont utilisés pour l'estimation des agrégats (micro-indices – voir à ce propos l'annexe C et en particulier la discussion du paragraphe C.3). Pour cette raison, le principe consistant à calculer un seul agrégat par territoire DOM est maintenu en base 2015.

Il est utile de procéder à une optimisation de l'échantillon conformément à ce qui est indiqué au paragraphe 4. Cette optimisation a été réalisée sur les données de l'IPC de 2013. La précision d'un indice calculé sur le champ de la collecte terrain est indiqué, par territoire, dans la table 7.

TABLE 7 – Précison des indices IPC en métropole et dans les DOM

Territoire	Nobs	σ	
	base 1998	IPC base 1998	IPC optimisé
métropole	112 767	0,14	0,022
Guadeloupe	5 666	0,63	0,10
Martinique	5 724	0,67	0,12
Guyane	4 206	0,64	0,10
Réunion	5 826	0,70	0,09

Note : écart-type, en points d'indice de décembre de l'année A, base=100 décembre A-1, pour un indice de prix fondé sur le champ restreint à la collecte terrain, hors produits frais. Le calcul correspond à la seule variance intra (ajouter 0,01 d'écart-type inter en métropole). L'optimisation correspond au calcul réalisé pour la base 2015, conformément aux principes exposés aux paragraphes 4, 6 et 7 ; le seul paragraphe 4 s'agissant des DOM. Cette évaluation ne prend pas en compte le phénomène d'arrondi des nombres réels de relevés issu de l'optimisation en nombres entiers, avec saturation tel qu'évoqué au paragraphe G.3 et la variance qui en découle (voir §C.4).

En optimisant le nombre de relevés variété par variété, sous contrainte que le coût de collecte d'ensemble, pour chaque territoire, reste identique à son niveau actuel, il serait possible d'atteindre le niveau de précision indiqué dans la table 7.

F L'apport des enquêtes Budget des familles et points de vente à la connaissance de la géographie de la dépense de consommation des ménages

Les pondérations spatiales utilisées dans l'IPC reposent sur une évaluation du poids des agglomérations dans la dépense de consommation des ménages. En pratique, les enquêteurs vont observer les prix dans les points de vente de l'agglomération. Par conséquent, l'univers du sondage lié aux points de vente est celui de la géographie

des lieux d'achat, et non celle de résidence des ménages. Il paraît dés lors cohérent de tenter d'évaluer le poids de dépense de consommation au lieu d'achat. Une première approche, utilisée jusqu'à présent dans l'IPC, consiste à évaluer ce poids proportionnellement à la démographie des ménages au lieu de résidence. Mais il est possible, en utilisant les enquêtes "Budget des familles" (2011) et "Points de vente" (2009) de : (1) améliorer l'évaluation la dépense de consommation au lieu d'achat par rapport à l'utilisation d'une *proxy* démographique et (2) de disposer des éléments de structure de répartition des achats par forme de vente. C'est l'objet des travaux présentés dans cette section.

Différents travaux ont été conduits sur les carnets de l'enquête budget des familles (2005 et 2011) et sur l'enquête points de vente 2009 de la division commerce, sur le champ de la France métropolitaine. Ils aboutissent à des chiffrages utiles pour l'établissement des pondérations de l'IPC base 2015 des groupes de produits et des agglomérations en France métropolitaine.
Pour cela plusieurs exploitations ont été faites au niveau France métropolitaine :

- calcul des dépenses faites par les ménages selon le type de produit (classées en 12 regroupements COICOP) et la taille d'unité urbaine du lieu d'achat : à partir des carnets de dépenses des enquêtes Budget des familles de 2005 et de 2011 ;
- calcul du chiffre d'affaires des établissements relevant du commerce de détail en magasin et de l'artisanat commercial (NAF 10.13B, 10.71B, 10.71C, 10.71D et division 47 hors groupe 47.8 et 47.9) en France métropolitaine selon la taille d'unité urbaine et selon la forme de vente à partir de l'enquête Points de vente 2009.

F.1 Calcul de dépenses au lieu d'achat à partir des carnets de dépenses issus des enquêtes Budget des Familles de 2005 et de 2011

L'enquête Budget des familles recense les dépenses de consommation par l'intermédiaire d'un questionnaire rétrospectif et d'un carnet de compte, remplis par les individus du ménage âgés de plus de 14 ans. Les carnets de compte remplis par les enquêtés de plus de 14 ans enregistrent les dépenses quotidiennes, ainsi que les petites dépenses irrégulières. La personne interrogée doit noter sur son carnet toutes les dépenses qu'elle effectue durant 14 jours (pour l'enquête de 2005) ou 7 jours (pour l'enquête de 2011). Dans les carnets de dépenses des ménages, les personnes indiquent le lieu d'achat ; on dispose donc des libellés de communes d'achat et du code postal de la commune d'achat. A partir de ces libellés de communes et des codes postaux, on a pu ramener un code géographique de commune en face de chaque ligne de dépense, puis une taille d'unité urbaine. L'exercice a été fait, pour la France métropolitaine, sur les données de BDF 2005, puis sur les données de BDF 2011. Le nombre de lignes de dépenses exploitées est deux fois moins important en 2011 qu'en 2005, car le remplissage du carnet ne concernait plus qu'une semaine de dépenses en 2011 au lieu de deux en 2005 : 495 900 lignes pour 2011 contre 990 861 lignes pour 2005, sur le champ France métropolitaine et fonctions de consommation (codage COICOP – voir table 8) 01 à 12.

L'outil carnet collectant les dépenses sur 2 semaines pour 2005 et sur une semaine pour 2011 ne permet pas de comptabiliser toutes les dépenses des ménages. Les dépenses importantes ou périodiques sont collectées par des questions directes, avec des périodes de référence variables et appropriées à chaque catégorie de dépenses. Les dépenses relevées dans les carnets codés (avec une taille d'unité urbaine du lieu d'achat) ne représentent que 58% des dépenses des ménages en 2005 et 55% en 2011. On ne prend pas en compte le poste 99 (produit non identifié) et le poste 13 (hors champ de la dépense de consommation finale des ménages) pour se ramener à un champ le plus proche possible de celui de l'IPC (dépense effective).

Suivant le type de produit, le taux de couverture des carnets dans la dépense globale et le taux de couverture des carnets codés dans la dépense globale est plus ou moins important (table 8) : les carnets codés représentent plus de 90% de la consommation des ménages pour les regroupements 01 et 02 de la COICOP à deux positions (produits alimentaires et boissons non alcoolisées, boissons alcoolisées-tabacs et stupéfiants). Ces regroupements contribuent à hauteur d'environ 20% des dépenses des ménages en 2011. Sur cette année, les regroupements "03 : Articles d'habillement et articles chaussants" et "05 : Ameublement, équipement ménager en entretien courant de la maison" sont couverts par les carnets à hauteur de 70% à 80% de leur montant total. Les autres regroupements connaissent des taux de couvertures par les carnets permettant la localisation des dépenses au lieu d'achat beaucoup plus faibles (table 8).

TABLE 8 – Calcul de taux de couverture des carnets codés au lieu d'achat

COICOP 2 positions	poids dans la dépense totale		taux de couverture carnet codé	
	2005	2011	2005	2011
1	15,5%	16,2%	**0,92**	**0,93**
2	2,6%	2,9%	**0,91**	**0,88**
3	7,9%	4,9%	0,48	0,76
4	14,9%	15,5%	0,37	0,20
5	7,5%	6,2%	0,59	0,69
6	3,7%	1,8%	0,78	1,63
7	16,0%	17,3%	0,55	0,60
8	3,7%	3,3%	0,27	0,12
9	9,2%	9,6%	0,64	0,52
10	0,7%	0,7%	0,27	0,13
11	5,7%	7,2%	0,61	0,51
12	12,6%	14,5%	0,43	0,30
total	100,0%	100,0%	0,58	0,55

Source : Insee, enquêtes budget des familles 2005 et 2011 ; 01 : Produits alimentaires et boissons non alcoolisées , 02 : Boissons alcoolisées, tabac et stupéfiants, 03 : Articles d'habillement et articles chaussants, 04 : Logement, eau, électricité, gaz et autres combustibles, 05 : Ameublement, équipement ménager en entretien courant de la maison, 06 : Santé (en 2011 le taux de couverture du carnet dépasse 100%, car les dépenses comprennent la part sécurité sociale et la part complémentaire, alors que dans la table de consommation, c'est un reste à charge), 07 : Transports, 08 : Communications, 09 : Loisirs et culture, 10 : Enseignement, 11 : Hôtels, cafés, restaurants, 12 : Autres biens et services.

En pratique, on adopte la géographie de la dépense au lieu d'achat fournie par la localisation des carnets au point de vente pour les regroupements COICOP de l'alimentaire et des boissons (01 et 02). Ce faisant, la géographie obtenue n'est pas pleinement représentative de la géographie de la dépense de consommation en points de vente physiques, mais ce choix résulte du compromis consistant à utiliser une information de qualité statistique satisfaisante et couvrant un champ de la consommation aussi large que possible. La méthode retenue pour le calcul des pondérations d'agglomérations consiste :

1. à calculer la répartition des dépenses de consommation pour le total 01+02 par taille d'unité urbaine conformément au tableau 9 ;

2. puis, au sein des classes de taille d'unité urbaine, répartir la dépense correspondante entre les unités urbaines de la classe proportionnellement au poids démographique (au sens du RP 2010) de chaque unité urbaine.

TABLE 9 – Répartition de la dépense de consommation et de la démographie par taille d'agglomérations

Taille de l'unité urbaine (UU de 2010)	poids démographique	poids dépenses au lieu d'achat (carnet BDF)		poids dépenses '01'+'02' au lieu d'achat (carnet BDF)		2011/2005	2011/2005 ('01'+'02')
France METRO	2007	2005	2011	2005	2011		
0-Rural	22,5	9,0	7,1	8,5	7,4	0,79	0,87
1-UU de 2 000 à 4 999 h	6,7	6,4	6,1	7,8	7,6	0,95	0,98
2-UU de 5 000 à 9 999 h	5,7	6,3	7,0	8,2	9,1	1,10	1,11
3-UU de 10 000 à 19 999 h	5,0	6,6	5,8	8,3	6,7	0,89	0,81
4-UU de 20 000 à 49 999 h	6,3	8,1	9,1	8,9	9,8	1,12	1,10
5-UU de 50 000 à 99 999 h	7,3	9,3	8,5	9,0	8,8	0,92	0,98
6-UU de 100 000 à 199 999 h	5,3	6,5	8,6	5,7	7,0	1,32	1,22
7-UU de 200 000 à 1 999 999 h	24,6	28,5	29,3	26,4	25,6	1,03	0,97
8-Agglomération de Paris	16,7	19,3	18,6	17,3	18,1	0,96	1,04
ensemble (France métropolitaine)	100,0	100,0	100,0	100,0	100,0		

Taille de l'unité urbaine (UU de 2010) - taille agglo IPC	poids démographique	poids dépenses au lieu d'achat (carnet BDF)		poids dépenses '01'+'02' au lieu d'achat (carnet		2011/2005	2011/2005 ('01'+'02')
France METRO	2007	2005	2011	2005	2011		
Rural (pas dans ipc)	22,5	9,0	7,1	8,5	7,4	0,79	0,87
D - UU de 2 000 à 19 999 h	17,4	19,3	18,9	24,3	23,4	0,98	0,96
C - UU de 20 000 à 99 999 h	13,6	17,4	17,6	17,9	18,6	1,01	1,04
B - UU de 100 000 h ou plus	29,9	35,0	37,9	32,1	32,6	1,08	1,02
A - Agglomération de Paris	16,7	19,3	18,6	17,3	18,1	0,96	1,04
ensemble	100,0	100,0	100,0	100,0	100,0		

01**** PRODUITS ALIMENTAIRES ET BOISSONS NON ALCOOLISE
02**** BOISSONS ALCOOLISEES, TABACS ET STUPEFIANTS

Source : Insee, enquêtes budget des familles 2005 et 2011

Le champ actuel de l'indice des prix porte sur l'ensemble des agglomérations de plus de 2000 habitants. L'étude menée à partir des enquêtes budgets des familles permet d'estimer le défaut de couverture géographique à 7% environ. En effet, alors que les habitants des communes du rural représentent 22,5% de la population métropolitaine (population municipale de 2007), 7% de la consommation totale est achetée dans le rural en 2011. Si on considère uniquement les regroupements 01 et 02, bien couverts, seulement 7,4% des dépenses sont faites dans un lieu d'achat situé dans le rural, en 2011.

F.2 Exploitation de l'enquête Points de vente 2009

L'enquête Points de vente 2009 a pour objectif de collecter le siret, le chiffre d'affaires, le nombre de personnes occupées et la surface de vente au niveau des établissements commerciaux. Pour ce faire, l'enquête interroge des unités légales du commerce de détail en magasin et de l'artisanat commercial qui répondent pour l'ensemble de ces magasins (NAF 10.13B, 10.71B, 10.71C, 10.71D et division 47 hors groupe 47.8 et 47.9) en France métropolitaine. Dans cette enquête, on se restreint aux unités légales et aux établissements marchands, actifs ou présumés actifs, exploitants, créés strictement avant le 2 janvier 2009, localisés en France métropolitaine et relevant du commerce de détail en magasin. L'échantillon a été stratifié selon trois modalités du zonage en aires urbaines (ZAUER 1999) :

- pôle urbain {1}
- commune mono ou multi-polarisée {2,3}
- espace à dominante rurale {4,5,6}

Ni le code géographique de la commune de localisation de l'établissement, ni la taille d'unité urbaine correspondante ne sont disponibles dans les données de l'enquête. À partir du code Siret et par appariement avec les

données par établissement, issues de Sirene, on dispose du code géographique et de la taille d'unité urbaine de l'établissement. À partir de ces données, on reconstitue un équivalent des formes de vente enquêtées dans l'IPC (table 10) et on calcule une répartition du chiffre d'affaire des points de vente d'une taille d'unité urbaine donnée selon les formes de vente reconstituées (table 11).

G La détermination des échantillons 2015, 2016 et 2017 de l'IPC

G.1 Le calcul du tableau d'équilibrage

On procède par optimisation. On cherche à déterminer le nombre de produits à créer et supprimer pour chaque varagglo de la base 2015 (unités urbaines). On se situe dans une année courante (2014, 2015, 2016) et l'échantillon à déterminer est celui de l'année suivante (resp. 2015, 2016, 2017).

Notations

- x_{av} : nombre de relevés à réaliser pour l'agglo a et la variété v, inconnue ;
- n^0_{av} : nombre de relevés initiaux correspondant au réalisé base 2015 figurant dans l'échantillon de l'année courante considérée (table PRODUIT des applications prix) ;
- N_a : moyens fixés par le site prix pour l'agglo a, en termes de nombre de relevés un mois courant de l'année suivante ;
- N_v nombre de relevés par variété (fixé par les équipes sectorielles de la division des prix à la consommation conformément aux principes indiqués ci-dessous) ;
- agglos : $a \in \{1, \ldots, A\}$, variétés : $v \in \{1, \ldots, V\}$.

On cherche à déterminer le nombre de relevés qu'il convient de réaliser dans chaque cellule varagglo (a, v) de sorte à minimiser l'écart à l'existant (pour éviter les modifications de collecte trop importantes et partant les créations[50] de produits trop nombreuses), sous contraintes de respect des marges portant sur le nombre total de relevés par variétés d'une part, et par agglos d'autre part. Mathématiquement, on cherche le vecteur $\mathbf{x}^\star_{av}$ tel que :

$$\mathbf{x}^\star_{av} = \underset{\mathbf{x}_{av}}{\operatorname{argmin}} \left\{ \sum_a \sum_v \left(x_{av} - n^0_{av}\right)^2 \middle| \mathcal{S} \right\}$$

où $\mathcal{S}$ est l'ensemble suivant :

$$\mathcal{S} = \left\{ \mathbf{x}_{av} \middle| \sum_a x_{av} = N_v \ ; \ \sum_v x_{av} = N_a \ ; \ J.\mathbf{x}_{av} = \boldsymbol{\rho} \right\}$$

où J est une matrice qui permet de fixer des contraintes d'égalité de certains x_{av} à une valeur $\boldsymbol{\rho}_{av}$ fixée *a priori* pour les couples (a, v) concernés. On suppose qu'il y a j contraintes égalité de ce type. Sur chaque ligne de J, toutes les composantes sont nulles sauf une qui est égale à 1.

On montre que la solution $X^\star$ ($\equiv \mathbf{x}^\star_{av}$) du programme précédent est la solution du système matriciel suivant[51] :

$$\left(\begin{array}{c|c} I & G' \\ \hline G & 0 \end{array}\right) \left(\begin{array}{c} X^\star \\ \hline \lambda \end{array}\right) = \left(\begin{array}{c} \nu \\ \hline \kappa \end{array}\right)$$

50. appelées OR ou "ordres de recherche" car se traduisant, dans le processus IPC, par des demandes émises en direction des enquêteurs de rechercher des produits sur le terrain.
51. La matrice G' désigne la transposée de la matrice G.

TABLE 10 – Les formes de vente et l'enquête "Points de vente" 2009

Codes FV	Signification	Exemple	Enquête PDV 2009
10	Hypermarché	Carrefour	Apet=4711F
20	Supermarché	Intermarché	Apet=4711D
25	Maxidiscount	Dia	
30	Supérette	Casino	Apet=4711C+4711B
40	Magasin populaire	Monoprix	Apet=4711E+4719B
50	Grand magasin	Printemps	Apet=4719A
60	Grande surface spécialisée	Picard, Décathlon	3 premières positions de Apet=474, 475, 476 + Apet=4711A, 4771Z, 4772A, 4772B, 4775Z, 4777Z, 4778A, 4773Z, 4774Z, 4776Z, 4778B, 4778C
70	Magasin traditionnel (-300m²)	Boulangerie, Zara	3 premières positions de Apet=472, 107 + Apet=1013B
80	Marché	Marché convention	
90	Services	Café, restaurant, coiffeurs	
99	autres	Services publics et administration, relevés en site prix, achat en bornes	

Source : Insee, enquête Points de vente 2009. À noter que la forme de vente Maxidiscount n'ayant pas pu être isolée, elle a été intégrée, dans cette analyse, à la forme de vente "Supermarché". Les commerces d'alimentation générale d'une surface de vente inférieure à 120 m^2 ont été regroupés avec les supérettes (qui ont une surface comprise entre 120 et 400 m^2). Les magasins populaires correspondent aux magasins multi-commerces (Apet=4711E). On y a aussi inclus le "non alimentaire non spécialisé - autres commerces (Apet=4719B)" qui sont des commerces de détail non spécialisé sans prédominance alimentaire en magasin d'une surface de vente inférieure à 2500 m^2. Quand ce type de commerce a une surface de vente égale ou supérieure à 2500 m^2, il est classé en "grands magasins". Les grandes surfaces spécialisées regroupent les commerces de détail de produits surgelés et commerces de détail dans les secteurs "équipement de l'information et de la communication", "équipement du foyer", "culture, loisirs, sport", "habillement, chaussures", "autres équipements de la personne", "produits pharmaceutiques, articles médicaux et orthopédiques", "autres commerces de détail". L'alimentation spécialisée et l'artisanat commercial sont classés en forme de vente "magasin traditionnel".

TABLE 11 – Répartition du chiffre d'affaires par taille d'unité urbaine et par forme de vente

Taille de l'unité urbaine (UU de 2010) France métro	Hypermarchés	Supermarchés	Supérettes	Magasins multi-commerces populaires	Grands magasins	Grandes surfaces spécialisées	magasins traditionnels	Ensemble
Rural (pas dans ipc)	3,0	33,6	6,8	0,5	0,1	39,9	16,2	100,0
D - UU de 2 000 à 19 999 h	23,3	27,6	1,7	0,8	0,2	37,9	8,5	100,0
C - UU de 20 000 à 99 999 h	35,5	18,1	1,1	1,2	0,2	37,1	6,8	100,0
B - UU de 100 000 h à 199 999 H	26,4	16,0	0,8	0,7	0,6	48,9	6,6	100,0
B - UU de 200 000 h ou plus	30,0	14,2	1,6	1,1	1,2	45,1	6,8	100,0
A - Agglomération de Paris	22,4	10,1	3,9	2,8	3,8	47,8	9,2	100,0

source : insee, enquête Points de vente 2009

où λ est un vecteur de multiplicateurs de Lagrange de dimension [52] $A + V + j$. La matrice I est l'identité de dimension $A \times V$. La matrice G est l'expression des $A + V + j$ contraintes définissant l'ensemble $\mathcal{S}$ et portant sur le vecteur X. Le vecteur ν correspond à l'empilement des n^0_{av} et le vecteur κ correspond, sur les V premières lignes, aux valeurs des N_v pour $v \in \{1, \ldots, V\}$, sur les A suivantes, aux valeurs de N_a pour $a \in \{1, \ldots, A\}$ et sur les j suivantes, à $\boldsymbol{\rho}$. En éliminant λ, on obtient pour $X^\star$ l'expression suivante [53] :

$$X^\star = \left(I - G'(GG')^- G\right)\nu + G'(GG')^- \kappa$$

En toute rigueur, le bon programme correspond au programme précédent complété de la positivité des paramètres inconnus, c'est-à-dire des composantes de X. Cette solution peut être obtenue par itération à partir du programme précédent en annulant progressivement les composantes de X négatives [54].

G.2 La pratique de l'optimisation

Au final, il apparaît avec les méthodes mises en œuvre et décrites ci-dessus, que le nombre d'ordres de recherche est fixé par le résultat de l'optimisation. En d'autres termes, il suffit de calculer, sur n'importe quel ensemble, les écarts positifs – respectivement négatifs – entre le nombre de relevés issu de l'optimisation $x^\star_{av}$ et le nombre de relevés initial n^0_{av} pour connaître le nombre d'ordres de recherche – respectivement de suppressions – que génère en moyenne la solution retenue.

Le levier des équipes sectorielles de la division des prix est le nombre total de relevés par variété. Ce nombre est introduit en *input* du calcul d'optimisation. Il *détermine* donc, avec les cibles par agglomérations fixées, pour l'année suivante, par les sites prix [55], le nombre d'ordres de recherche et de suppressions de produits que va générer la solution obtenue. Si on veut, par exemple, réduire le nombre d'ordres de recherche sans changer

52. Il y a autant de multiplicateurs que de contraintes.
53. La matrice G n'est pas de plein rang sur ses lignes car la dernière ligne est la somme de toutes les autres. Ceci traduit le fait que $\sum_v \sum_a x_{av} = \sum_a \sum_v x_{av}$. La matrice $(GG')^-$ désigne l'inverse généralisée de la matrice GG'.
54. Ceci revient à supprimer les lignes de X et de ν, ainsi que les colonnes de G correspondantes.
55. Dans la pratique, les cibles "base 2015" sont fixées globalement par agglo en sommant le nombre de relevés par variété qui résultant de l'optimisation présentée au §7.3. Une organisation a été mise en œuvre par la division des prix à la consommation et les sites prix pour atteindre en 3 ans, entre 2014 et 2017, les cibles de la base 2015 sur chaque agglo. Ainsi, en 2017, le volume de relevés sera, sur chaque agglo, conforme à celui déterminé par le calcul de variance présenté au §7.3, à l'exception des agglos pour lesquelles l'analyse terrain à fixé un nombre maximal de relevés atteignable inférieur à l'optimum calculé. Dans ce dernier cas, les cibles ont été fixées au maximum atteignable compte tenu des règles de nombre de relevé maximal par forme de vente. Rappelons que ces règles, établies de longue date dans l'IPC, sont destinées à limiter les "effets de grappes" qui résulteraient d'une trop grande concentration de relevés par points de vente. Rappelons enfin, à propos des cibles de nombre de relevés par varagglos, qu'il peut être procédé chaque année à une mise à jour du calcul d'optimisation du nombre de relevés et les cibles de nombre de relevés par agglo sont donc susceptibles d'être modifiées en conséquence (i.e. selon un rythme annuel).

les cibles globales d'activité du réseau (qui elles sont simultanément déterminées par la somme des cibles de relevés par agglo et la somme des cibles par variété), il faut faire en sorte qu'il y ait moins d'ordre de recherche et moins de suppressions. Et il faut que les ordres de recherche en moins soient compensés par des suppressions en moins grand nombre.

G.3 Passer de valeurs réelles du nombre de relevés à des valeurs entières

Quatre cas sont à distinguer selon le type de variété concerné :

1. celui des variétés homogènes (HO) dont on impose que le nombre de relevés par varagglo soit quelconque. Moyennant quoi, pour ces varagglos, une fois adopté un nombre de relevés égal à la partie entière d'une variable continue, le problème de l'arrondi revient à arrondir par 0 ou 1 relevé une variable continue – partie décimale complément de la partie entière précédente – comprise dans l'intervalle $[0, 1]$.
2. celui des variétés homogènes (HO) dont on impose que le nombre de relevés par varagglo soit égal à 0, 2 ou plus (pas de situation de varagglo avec un seul relevé).
3. celui des variétés hétérogènes ou biens durables (HE-BD) dont on impose que le nombre de relevés par varagglo soit supérieur ou égal à 4, ou bien nul. Moyennant quoi, pour ces varagglos, le problème de l'arrondi revient à arrondir par 0 ou 4 une variable continue comprise dans l'intervalle $[0, 4]$. Si la variable est supérieure à 4, on se retrouve dans le cas 1 ci-dessus.
4. celui des variétés hétérogènes ou biens durables (HE-BD) dont on impose que le nombre de relevés par varagglo soit supérieur ou égal à 4, ou bien nul et tienne compte de l'existant.

Pour toutes les var-agglos où, quel que soit le type de variété, le nombre de relevés optimisé est supérieur à la borne inférieure requise (1, 2, ou 4 relevés selon le cas), le nombre de relevés retenu est déterminé en sommant la partie entière du nombre de relevés optimisé et le résultat de la procédure d'arrondis 1 appliqué à la partie décimale (i.e. le complément à la partie entière) du nombre de relevés optimisé.

Les variétés HO des carburants et des journaux sont traitées selon la procédure 1. Toutes les autres variétés HO sont traitées selon la procédure 2.

S'agissant de variétés HE ou BD, deux situations peuvent survenir :

— Si la varagglo ne concerne pas une variété ou une agglo qui connaissent des variations très importantes de nombre de relevés, alors le nombre de relevé est déterminé conformément à la procédure 3.
— Sinon, il est déterminé conformément à la procédure 4.

Les procédures 1, 2, 3, 4 sont respectivement détaillées aux paragraphes G.3.1, G.3.2, G.3.3 et G.3.4.

Soit X_i une variable à arrondir. X_i est dans un intervalle $[0, 2]$ dans le cas général pour les variétés HO, $[0, 1]$ dans certains cas particuliers, et $[0, 4]$ pour les variétés HE-BD. La variable arrondie est Y_i et on veut que $\mathbb{E}(Y_i) = X_i$.

G.3.1 Le cas où $Y_i \in \{0, 1\}$

Dans le cas où $Y_i \in \{0, 1\}$, on applique la méthode de base suivante. On note $\mathbb{P}(Y_i = 1) = \pi_i$. Y_i est une variable de Bernouilli. On cherche π_i de sorte que $\mathbb{E}(Y_i) = X_i$. Puis,

$$\begin{aligned} \mathbb{E}(Y_i) &= 1 \times P(Y_i = 1) + 0 \times P(Y_i = 0) \\ &= \pi_i \end{aligned}$$

Par conséquent $\pi_i = X_i$.

Encadré : Le tirage À ce stade, on souhaite simuler une réalisation de Y_i, variable de Bernouilli $\mathscr{B}(\pi_i)$. Ceci s'obtient en réalisant une variable uniforme $u \hookrightarrow \mathscr{U}_{[0,1]}$ et

— si $u \leqslant \pi_i$, alors $Y_i = 1$

— si $u > \pi_i$, alors $Y_i = 0$

Ainsi, on obtient bien que $\mathbb{P}(Y_i = 1) = \pi_i$.

G.3.2 Le cas où $Y_i \in \{0, 2\}$

On procède comme précédemment avec $Y_i \in \{0, 2\}$. On note de même $\mathbb{P}(Y_i = 2) = \pi_i$. Puis,

$$\begin{aligned} \mathbb{E}(Y_i) &= 2 \times P(Y_i = 2) + 0 \times P(Y_i = 0) \\ &= 2 \times \pi_i \end{aligned}$$

Par conséquent $\pi_i = \frac{X_i}{2}$.
Le tirage est alors effectué (voir encadré "Le tirage" ci-dessus).

G.3.3 Le cas où $Y_i \in \{0, 4\}$, méthode de base

On procède comme précédemment avec $Y_i \in \{0, 4\}$. On note de même $\mathbb{P}(Y_i = 4) = \pi_i$. Puis,

$$\begin{aligned} \mathbb{E}(Y_i) &= 4 \times P(Y_i = 4) + 0 \times P(Y_i = 0) \\ &= 4 \times \pi_i \end{aligned}$$

Par conséquent $\pi_i = \frac{X_i}{4}$.
Le tirage est alors effectué (voir encadré "Le tirage" ci-dessus).

G.3.4 Le cas où $Y_i \in \{0, 4\}$, méthode avec une variable auxiliaire

Quand des relevés sont déjà effectués dans une agglomération au titre d'une variété donnée, la méthode d'arrondi présentée ci-dessus, par construction indépendante de l'existant, peut aboutir à supprimer des relevés là où des relevés existent et à l'opposé, à en créer là où il n'en existait pas. Il peut donc être intéressant, dans certains cas, de conditionner l'arrondi à une variable auxiliaire, comme par exemple le nombre de relevés existants ou une variable en dérivant. On suppose donc à présent que l'on dispose d'une variable auxiliaire Z_i dont on voudrait qu'elle soit prise en compte pour "orienter" le tirage de Y_i. Pour l'exemple, on se place dans le cas où $Y_i \in \{0, 4\}$.

Pour être plus précis, *on souhaiterait que, conditionnellement à Z_i, plus Z_i est grand, plus la probabilité que Y_i soit égale à 4 est élevée.* On s'intéresse donc à la probabilité conditionnelle $\mathbb{P}(Y_i|Z_i)$. Par la relation de la probabilité totale et si on suppose que Z_i prend ses valeurs discrètes dans un ensemble J, alors :

$$\mathbb{P}(Y_i) = \sum_{j \in J} P(Y_i|Z_i = j) \times \mathbb{P}(Z_i = j)$$

La question est alors : comment choisir les $\mathbb{P}(Z_i = j)$ pour que les propriétés évoquées plus haut (en italique) soient vérifiées ?

Pour fixer les idées, on suppose que Z_i se fonde sur une variable $\tilde{Z}_i$, par exemple un nombre de relevés existant et les modalités de Z_i sont fixées, par rapport à $\tilde{Z}_i$, de la manière suivante (Z_i prend deux modalités) :

$$Z_i \ : \ \begin{cases} \tilde{Z}_i \geqslant 2 \\ \tilde{Z}_i < 2 \end{cases}$$

On note :

$$\left\{\begin{array}{rcl}\mathbb{P}(Y_i=4|\tilde{Z}_i\geqslant 2) & = & \alpha_i\\ \mathbb{P}(Y_i=4|\tilde{Z}_i<2) & = & \beta_i\end{array}\right.$$

On a choisi de considérer que la probabilité conditionnelle est une fonction croissante de Z_i, donc $\alpha_i > \beta_i$. Par construction,

$$\left\{\begin{array}{rcl}\mathbb{P}(Y_i=0|\tilde{Z}_i\geqslant 2) & = & 1-\alpha_i\\ \mathbb{P}(Y_i=0|\tilde{Z}_i<2) & = & 1-\beta_i\end{array}\right.$$

Puis, la formule de la probabilité totale donne :

$$\mathbb{P}(Y_i=4)=\mathbb{P}(\tilde{Z}_i\geqslant 2)\times\alpha_i+\mathbb{P}(\tilde{Z}_i<2)\times\beta_i$$

et

$$\mathbb{P}(Y_i=0)=\mathbb{P}(\tilde{Z}_i\geqslant 2)\times(1-\alpha_i)+\mathbb{P}(\tilde{Z}_i<2)\times(1-\beta_i)$$

Enfin, on souhaite que la probabilité $\mathbb{P}(Y_i=4|\tilde{Z}_i\geqslant 2)$ soit la plus grande possible.

Pour résumer, on cherche α_i et β_i tels que :

$$\left\{\begin{array}{l}\alpha_i\mathbb{P}(\tilde{Z}_i\geqslant 2)+\beta_i\mathbb{P}(\tilde{Z}_i<2)=\frac{X_i}{4}\\ \mathbb{P}(Y_i=4|\tilde{Z}_i\geqslant 2)=\alpha_i \text{ est maximal}\end{array}\right. \tag{27}$$

Par hypothèse, $\alpha_i\in[0,1]$, de même que β_i. $\mathbb{P}(\tilde{Z}_i<2)=1-\mathbb{P}(\tilde{Z}_i\geqslant 2)$. On note[56] $\gamma_i=\mathbb{P}(\tilde{Z}_i\geqslant 2)$. La première relation du système (27) s'écrit :

$$\alpha_i\gamma_i+\beta_i(1-\gamma_i)=\frac{X_i}{4}$$

Soit,

$$\alpha_i=\frac{1}{\gamma_i}\left[\frac{X_i}{4}-\beta_i(1-\gamma_i)\right]$$

α_i apparaît comme une fonction décroissante de β_i (on rappelle que $\beta_i\in[0,1]$). Le tableau de variation de la fonction est le suivant :

β_i	0		1
α_i	$\frac{1}{\gamma_i}\frac{X_i}{4}$	$\searrow$	$\frac{1}{\gamma_i}\frac{X_i}{4}-\frac{1-\gamma_i}{\gamma_i}$

On peut donc choisir β_i de sorte que α_i soit maximale. Cependant, α_i doit être dans l'intervalle $[0,1]$. Son maximum est donc $\min\left\{1,\frac{1}{\gamma_i}\frac{X_i}{4}\right\}$. Il convient donc de discuter en fonction de la position de $\frac{1}{\gamma_i}\frac{X_i}{4}$ par rapport à 1. Deux cas peuvent se présenter :

1. Si $\frac{1}{\gamma_i}\frac{X_i}{4}\leqslant 1$ alors $\alpha_i^\star=\frac{1}{\gamma_i}\frac{X_i}{4}$ est la solution recherchée. Moyennant quoi, $\beta_i^\star=0$.
2. Si $\frac{1}{\gamma_i}\frac{X_i}{4}>1$ alors la solution consiste à rechercher le β_i tel que $\alpha_i=1$. Ce β_i vaut :

$$\beta_i^\star=\frac{1}{1-\gamma_i}\times\left(\frac{X_i}{4}-\gamma_i\right)$$

   Moyennant quoi, $\alpha_i^\star=1$.

56. γ_i ainsi défini peut être estimé empiriquement (voir *infra*).

Remarque : on peut vérifier au passage que $\mathbb{P}(Y_i = 4) = \frac{X_i}{4}$. En effet,

$$\mathbb{P}(Y_i = 4) = \gamma_i \times \alpha_i^\star + (1 - \gamma_i) \times \beta_i^\star$$

En particulier,

— Si $\frac{1}{\gamma_i}\frac{X_i}{4} \leqslant 1$ alors $\mathbb{P}(Y_i = 4) = \gamma_i \times \left(\frac{1}{\gamma_i}\frac{X_i}{4}\right) + (1 - \gamma_i) \times 0 = \frac{X_i}{4}$.

— Si $\frac{1}{\gamma_i}\frac{X_i}{4} > 1$, alors $\mathbb{P}(Y_i = 4) = \gamma_i \times 1 + (1 - \gamma_i) \times \frac{1}{1-\gamma_i} \times \left(\frac{X_i}{4} - \gamma_i\right) = \frac{X_i}{4}$

Mise en œuvre : Pour un ensemble d'observations [57] $\mathscr{E}$, on détermine empiriquement $\hat{\gamma}_i = \mathbb{P}(\tilde{Z}_i \geqslant 2)$. Puis pour chacune des observations i de l'ensemble $\mathscr{E}$, on effectue les opérations suivantes :

1. On détermine $(\alpha_i^\star, \beta_i^\star)$ conformément à la procédure décrite ci-dessus.
2. Puis,
   — Si $\tilde{Z}_i \geqslant 2$, alors on retient $Y_i = 4$ avec une probabilité $\alpha_i^\star$ (loi de Bernouilli de paramètre $\alpha_i^\star$ – voir aussi l'encadré "Le tirage" ci-dessus).
   — Si $\tilde{Z}_i < 2$, on prend $Y_i = 4$ avec une probabilité $\beta_i^\star$ (loi de Bernouilli de paramètre $\beta_i^\star$ – voir aussi l'encadré "Le tirage" ci-dessus).

57. D'un point de vue pratique, $\mathscr{E}$ est l'ensemble des varagglos existantes concernées par le cas traité dans la présente section G.3.4.