

# Écartspatiaux de niveaux de prix entre régions et villes françaises avec des données de caisse

## *Spatial Differences in Price Levels between French Regions and Cities with Scanner Data*

Isabelle Léonard\*, Patrick Sillard\*, Gaëtan Varlet\*  
et Jean-Paul Zoyem\*

**Résumé** – Cette étude s’appuie sur les données de caisse de la grande distribution transmises quotidiennement à l’Insee en 2013. Elle vise à calculer des indices mesurant des différences de niveau de prix à la consommation entre territoires métropolitains dans l’alimentaire vendu en supermarché. Un indice hédonique fondé sur la régression du prix du produit sur des indicatrices de codes-barres et de territoires est développé. Plusieurs évaluations sont déterminées sur différentes semaines, une semaine de données permettant déjà d’atteindre une précision considérable. La dispersion des niveaux de prix entre régions ou entre grandes agglomérations est limitée et, pour l’essentiel, robuste au choix de la semaine : les prix les plus élevés sont observés en région parisienne ainsi qu’en Corse, les écarts étant de l’ordre de quelques points de pourcentage. Le rapprochement de ces résultats nouveaux et des travaux réalisés par l’Insee entre 1970 et 2000 montre que les écarts de prix dans l’alimentaire entre territoires métropolitains sont essentiellement structurels et évoluent peu dans le temps.

**Abstract** – This study is based on scanner data from large retailers sent daily to Insee in 2013. Its aim is to compute indices that measure differences in consumer price levels between different areas of metropolitan France, focusing specifically on food products sold in supermarkets. A hedonic index based on the regression of the product price on barcode and territory dummies is developed. Several assessments are carried out over different weeks, with one week of data already providing a great degree of accuracy. The dispersion of price levels between regions or large conurbations is limited and, for the most part, robust to the choice of week. The highest prices are found in the Paris region and Corsica, with a magnitude of differences in the order of a few percentage points. A comparison of the new findings with research conducted by Insee between 1970 and 2000 shows that differences in food prices across areas of metropolitan France are essentially structural and change little over time.

Codes JEL / JEL Classification : E31, C8, D1

Mots-clés : niveaux de prix, comparaison spatiale, données de caisse

Keywords: price levels, spatial comparison, scanner data

Rappel :

Les jugements et opinions exprimés par les auteurs n’engagent qu’eux mêmes, et non les institutions auxquelles ils appartiennent, ni a fortiori l’Insee.

\* Insee (isabelle.leonard@insee.fr ; patrick.sillard@insee.fr ; gaetan.varlet@insee.fr ; jean-paul.zoyem@insee.fr)

Reçu le 18 septembre 2017, accepté après révisions le 17 janvier 2019

Pour citer cet article : Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2019). Spatial Differences in Price Levels between French Regions and Cities with Scanner Data. *Economie et Statistique / Economics and Statistics*, 509, 69–82. <https://doi.org/10.24187/ecostat.2019.509.1983>

Le système d'observation des prix à la consommation mis en place par la statistique publique française vise essentiellement à déterminer les écarts temporels de prix, c'est-à-dire l'inflation. L'indice des prix à la consommation (IPC) est une mesure de cette grandeur. Pour cela, mois après mois, les enquêteurs de l'Insee retournent dans les mêmes points de vente observer le prix des mêmes produits, et c'est à partir des évolutions élémentaires de prix observées au niveau de chaque produit suivi dans l'IPC qu'est calculée l'évolution moyenne d'ensemble. L'intuition suggère que les prix collectés pour les besoins de l'IPC pourraient aussi permettre de déterminer des écarts moyens de niveaux de prix entre différents territoires d'intérêt. Pourtant ce n'est pas en général le cas. En effet, l'enjeu, pour mesurer une évolution moyenne des prix, est d'assurer qu'entre deux périodes comparées, on compare effectivement les mêmes produits. Dans le même esprit, comparer des niveaux de prix entre territoires suppose d'observer les prix de produits identiques sur les territoires dont les niveaux de prix sont comparés. Comme ce dernier enjeu, spécifique à la comparaison de niveaux territoriaux, n'en est pas un pour l'IPC, l'identification des produits réalisée pour l'IPC n'est en général pas suffisamment riche pour assurer que deux produits observés dans deux points de vente différents sont identiques. Par ailleurs, l'échantillon de produits suivis dans l'IPC est obtenu par sondage et optimisé pour atteindre une précision satisfaisante de la mesure d'inflation au plan national. Descendre plus finement en localisation géographique se heurte mécaniquement à la faiblesse du nombre de relevés effectués sur des territoires d'extension limitée. Au final, même si les produits étaient mieux identifiés dans l'IPC, on ne pourrait que difficilement effectuer des comparaisons satisfaisantes de niveaux de prix entre territoires.

À l'opposé, les données de caisse ne comportent pas certaines des limitations évoquées au sujet des données de l'IPC pour déterminer des écarts de niveaux de prix spatiaux : 1) le code-barres (appelé aussi EAN pour *European Article Number*) est un identifiant univoque du produit<sup>1</sup> ; 2) les données de caisse couvrent l'exhaustivité des transactions relatives aux produits de l'alimentaire industriel<sup>2</sup> hors produits frais – i.e. les fruits, légumes, crustacés ainsi qu'une partie des poissons et des viandes –, aux boissons alcoolisées ou non et à une partie des biens manufacturés vendus dans les hypermarchés et supermarchés de France métropolitaine. La première propriété précédente assure que la comparaison de prix d'un même code-barres vendu dans deux magasins

différents conduit à comparer *de facto* le même produit. La seconde propriété assure que les échantillons disponibles sont suffisamment vastes pour permettre une comparaison à des niveaux de détail fins.

L'Insee a engagé une expérience pilote destinée à intégrer progressivement les données de caisse dans la détermination de l'IPC. Pour cela, l'Insee reçoit quotidiennement depuis la fin de l'année 2012 les données de caisse de plusieurs groupes de la grande distribution. Les groupes participant à l'expérience pilote représentent environ 30 % du champ potentiel, c'est-à-dire celui qui correspondrait aux transactions quotidiennes de l'ensemble des enseignes de la grande distribution présentes sur le territoire métropolitain. Les données de caisse regroupent, pour chaque magasin, la liste des transactions du jour, c'est-à-dire la liste des codes-barres vendus, ainsi que les quantités écoulées et les prix de vente<sup>3</sup> correspondants.

Un des avantages essentiels des données de caisse est la richesse d'information. L'extraordinaire volume de données permet d'envisager de fournir un niveau de détail d'information sur la connaissance des niveaux de prix inatteignable avec le dispositif habituel de collecte. En outre, ces données comportent simultanément l'information de prix et celle des quantités de produits vendus, ce qui constitue une matière inédite pour la statistique de prix puisque celle-ci se fonde, habituellement, sur la seule connaissance des prix de détail. Si les premières applications portent naturellement sur la détermination de l'inflation dans les pays (voir par exemple Reinsdorf, 1999 ; de Haan & van der Grient, 2011), d'autres applications statistiques sont possibles. La comparaison de niveaux de prix entre pays reste compliquée car les produits élémentaires, le système de codage des produits ou simplement les systèmes d'information des enseignes ne sont généralement pas suffisamment homogènes pour permettre un rapprochement en masse des EAN. En revanche, pour un pays donné, lorsque le système d'information des données de caisse fournit en même temps des informations détaillées selon les lieux d'achat, il devient possible d'utiliser les données de caisse pour déterminer des écarts de niveau de prix entre zones géographiques. C'est cette

1. C'est-à-dire que deux produits différents (vus comme tels par le consommateur) ont nécessairement deux EAN différents. En revanche, deux EAN différents peuvent désigner le même produit.

2. Sauf mention contraire, on entend, dans cet article, le champ de l'alimentaire industriel comme étant celui des produits d'alimentation, hors produits frais (i.e. fruits et légumes frais, crustacés ainsi qu'une partie des poissons et des viandes), et des boissons alcoolisées ou non, vendus en supermarchés (voir la section sur les données pour plus de détails).

3. Parfois le chiffre d'affaires correspondant, plutôt que le prix.

question qui est examinée ici, pour le champ de l'alimentaire industriel, à l'aide d'un jeu de données de caisse dont l'Insee a pu disposer sur l'année 2013.

La comparaison spatiale de niveaux de prix est une pratique courante dans de nombreux pays, généralement coordonnée par des institutions internationales. S'agissant d'indices bilatéraux, l'opération consiste à définir des classes de produits homologues entre les pays, à déterminer une structure de consommation en dépense pour le couple de pays considérés, à identifier des produits représentatifs des consommations nationales et comparables dans leur usage entre les deux pays, puis à calculer un indice bilatéral caractérisant la différence de niveaux de prix entre pays. Une des grandes difficultés de ce genre d'opération est de déterminer des classes de produits réellement homologues, c'est-à-dire qui correspondent à un « usage » équivalent dans les différents pays comparés. En effet, faute d'être capable d'identifier des produits identiques – lesquels n'existent d'ailleurs pas toujours, en particulier lorsque les pays sont assez différents dans leurs cultures et niveaux de vie –, les institutions qui coordonnent ces comparaisons fondent la mesure des différences de prix sur le rapprochement de produits dont les caractéristiques sont aussi proches que possible. Si cette approche donne une bonne approximation des écarts de niveaux de prix, fondée sur un compromis entre définition des produits et comparabilité, elle reste cependant discutable du fait de ce compromis. Les limites de ces comparaisons dites en « parité de pouvoir d'achat » sont bien connues et détaillées dans la littérature (voir par exemple Deaton & Heston, 2010). Il faut retenir de cette littérature que les discussions portent sur deux points, de nature différente mais qui, tous deux, sont d'importance limitée pour l'exercice de comparaison conduit ici sur des données de caisse et pour comparer des régions françaises entre elles. Le premier point de débat concerne l'exercice de rapprochement des produits qui devient potentiellement impossible lorsque les zones comparées sont très différentes ; dans le cas qui nous intéresse ici, les zones comparées – des régions de France métropolitaine ou des agglomérations – sont très homogènes dans leurs usages de consommation. Le second point porte sur la méthode de calcul des indices d'écart de niveaux. En pratique, les méthodes utilisées conduisent à des indices qui diffèrent d'autant moins que les prix et structure de consommation sont proches entre les zones comparées.

Potentiellement plus critique est la question du champ de la comparaison. Par construction, les

résultats présentés dans cet article portent sur le champ sur lequel on dispose de données de caisse. D'une part, il s'agit en l'occurrence du champ des produits de l'alimentaire (hors produits frais) et des boissons alcoolisées ou non, vendus en grandes surfaces (i.e. alimentaire industriel). Les achats alimentaires réalisés dans d'autres types de points de vente ne sont pas inclus et les résultats obtenus ne sont donc pas représentatifs de l'ensemble de la consommation alimentaire. D'autre part, en 2013, l'Insee ne disposait des données de caisse que de quelques enseignes de la grande distribution. Les ventes correspondantes représentaient environ 30 % des ventes de la grande distribution, dans l'alimentaire industriel, en métropole. Par conséquent, il est possible que les comparaisons régionales de niveau de prix examinées dans cet article soient affectées d'un biais lié à cette sélection d'enseignes. La section consacrée à la présentation des données examine plus en détail ces questions de couverture et montre, en particulier, que la structure de consommation obtenue à partir du champ restreint est conforme à la répartition géographique de la population française. L'impact éventuel de la politique géographique de prix des enseignes sélectionnées dans l'échantillon est plus compliqué à cerner : si la politique est spécifique à l'enseigne et que, dans le même temps, le poids de cette enseigne dans le territoire comparé diffère entre l'échantillon de l'Insee et la situation générale toutes enseignes réunies, alors l'indice du territoire estimé sur fondement de l'échantillon particulier sera différent de celui obtenu toutes enseignes confondues. *A priori*, les effets de la concurrence locale tendent toutefois à forcer à l'harmonisation des structures de prix entre enseignes et dans l'espace. Par conséquent, des estimations fondées sur un sous-échantillon couvrant 30 % de la population d'ensemble devraient, dans ce contexte, permettre de tirer des enseignements de portée relativement générale.

La suite de l'article est organisée ainsi : une première section présente les résultats d'autres exercices de mesure des écarts de prix entre régions métropolitaines et entre grandes agglomérations réalisées par l'Insee depuis 1971. Les résultats nouveaux obtenus à partir des données de caisse dans cette étude sont ainsi mis en perspective de résultats comparables et plus anciens. Des statistiques descriptives sont données dans la deuxième section, puis la section suivante présente la modélisation retenue pour l'analyse des données. Une dernière section présente les différents résultats obtenus et une analyse de robustesse, qui reprend les différents axes de discussion évoqués plus haut.

## Quelques expériences passées de comparaisons spatiales au niveau du territoire métropolitain

Les travaux de comparaison de niveaux de prix entre régions métropolitaines sont très anciens puisque figurent, dans les publications de la Statistique Générale de la France (SGF), à la fin du 19<sup>e</sup> siècle et au début du 20<sup>e</sup>, des tableaux comparatifs de prix de détail moyens par denrées observés dans différentes villes de France. Il faut toutefois attendre des années plus récentes pour disposer de comparaisons qui couvrent un champ significatif de la consommation et qui sont fondées sur un nombre important de produits. Techniquement, ces travaux<sup>4</sup> consistent, s'agissant des comparaisons de niveaux de prix métropolitains, à calculer un ratio de prix moyen entre le territoire considéré et la France entière, pour des produits représentatifs de la consommation d'une variété de produits donnée, puis à agréger les différences ainsi mesurées au niveau des variétés de produits en une moyenne pondérée nationale. La pondération appliquée pour cette moyenne correspond à la structure de consommation nationale, sans particularités locales, au motif que les structures locales de consommations sont très peu différentes de la structure nationale (Mineau, 1987 ; Anxionnaz & Mothe, 2000). Parmi les travaux plus récents que ceux de la SGF, on peut citer Piccard (1972) et Baraille (1978) qui traitent des différences de niveaux entre villes métropolitaines. Les résultats de ces deux études sont repris dans le tableau 1. Ces deux études concluent, dans le même sens, que dans le champ de l'alimentaire et des boissons, l'agglomération parisienne et la Corse sont les lieux où les prix de l'alimentaire et des boissons (alcoolisées ou non) sont les plus élevés de métropole. Par ailleurs, elles mettent en évidence une dispersion somme toute peu importante, comprise dans une fourchette un peu inférieure à 10 points de pourcentage<sup>5</sup>.

Ces travaux ont été complétés par Baraille & Bobin (1981) en utilisant une analyse par type de territoire basée sur une nouvelle enquête réalisée en 1981. Ce type d'analyse faisait écho à des résultats du même type obtenus par Piccard (1972).

Plus récemment, Mineau (1987) donne la ventilation par grande agglomération des écarts de niveaux de prix dans l'alimentaire et les boissons pour l'année 1985 ; la Division prix de détail de l'Insee (1990) en fait de même pour l'année 1989. Ces deux groupes de résultats témoignent d'une stabilité des écarts en niveaux de prix entre les différentes agglomérations que l'on peut constater au tableau 1. Naturellement, les deux années

étudiées (ici 1985 et 1989) sont proches mais le constat est similaire avec l'année 1977 pourtant plus éloignée (tableau 1). À nouveau, dans ces travaux, on constate que les prix dans l'alimentaire et des boissons sont plus élevés en Corse que partout ailleurs. Puis, vient l'agglomération parisienne dans laquelle les prix à la consommation sont de 2 à 3 % plus élevés que dans les autres villes de province.

L'étude sur 1995 menée par Guglielmetti (1996) établit quant à elle que l'écart de niveau moyen des prix de l'alimentaire et des boissons (alcoolisées ou non, y. c. tabac) en Corse est sensiblement plus élevé qu'en 1989, puisqu'il atteint 8.5 % par rapport à Paris, l'écart entre Paris et Marseille s'étant maintenu sur la période.

Les derniers travaux menés, plus globaux, n'indiquent pas de modifications notables par rapport à ces constats. Fesseau *et al.* (2008) constatent que les prix de l'alimentaire et des boissons non alcoolisées sont environ 5.7 % plus élevés en Île-de-France qu'en province sur l'année 2006. Nicolaï (2010) établit, à partir de l'enquête de comparaison spatiale des niveaux de prix menée par l'Insee en 2010, que les niveaux moyens des prix de l'alimentaire et des boissons non alcoolisées sont environ 8.6 % plus élevés en Corse que sur le continent pris dans son ensemble. Enfin, la réédition de cette enquête en 2015 a permis de montrer que les prix de l'alimentaire<sup>6</sup> et des boissons non alcoolisées, cette année-là, étaient de 6.5 % plus élevés en région parisienne qu'en province et de 2.1 % plus élevés en Corse qu'en région parisienne (Clé *et al.*, 2016). Ces derniers résultats, établis à partir de données collectées pour mesurer les écarts de niveaux de prix, confirment donc, dans le champ de l'alimentaire, la hiérarchie et les ordres de grandeurs établis précédemment.

Au final, ces différentes études dont le champ, la méthodologie et la nature diffèrent quelque peu, donnent des résultats globalement cohérents : les différences de niveaux de prix sont des

4. À l'exception notable des travaux les plus récents de comparaison spatiale de prix adossés à des enquêtes ad hoc (Nicolaï, 2010 ; Berthier *et al.*, 2010 ; Clé *et al.*, 2016). Ces travaux s'appuient sur une approche inspirée des enquêtes européennes harmonisées de mesure des parités de pouvoir d'achat et sont fondés sur des indices de prix de Fisher, calés sur des structures de consommations spécifiques à chacun des territoires comparés. Cette approche se justifie pleinement lorsque les structures de consommation diffèrent significativement entre les territoires comparés, comme c'est le cas entre les DOM et la métropole par exemple. En revanche pour des comparaisons de régions métropolitaines, les différences de structures régionales sont très limitées et l'enjeu de leur prise en compte est secondaire.

5. L'étude de Baraille (1978) mesurait un écart de 8 % entre les prix de l'alimentaire et des boissons dans l'agglomération où ils étaient les plus élevés (Ajaccio-Bastia) et les moins élevés (Angers).

6. Incluant également les produits frais.

Tableau 1  
**Écarts de prix moyens observés en France métropolitaine dans le champ de l'alimentaire et des boissons**

Territoire	Indice, selon les résultats de :				
	Piccard, 1972 pour 1971	Baraille, 1978 pour 1977	Mineau, 1987 pour 1985	Insee, 1990 pour 1989	Guglielmetti, 1996 pour 1995
Agg. parisienne	100	100.0	100.0	100.0	100.0
Lyon	100	97.5	99.0	98.7	
Marseille	104	98.3	99.5	97.5	97.0
Bordeaux	100	94.1	96.7	96.6	
Rennes	97	93.8	92.8	94.4	
Reims		97.2	97.7	97.8	
Rouen		97.7	95.9	95.1	
Strasbourg		98.1	97.0	98.2	
Lille		97.6	95.3	95.7	
Orléans		95.7	96.2	95.7	
Limoges		97.4	96.7	97.1	
Ajaccio-Bastia		100.5	105.1	103.6	108.5
Clermont-Ferrand		99.0	100.9	98.5	
Toulouse		95.1	98.5	98.9	
Dijon		96.7	96.9	97.9	
Nantes		93.6	93.7	94.7	
Nancy		95.0	98.9	97.1	
Poitiers		94.2	92.5	92.2	
Montpellier		96.4	100.1	100.4	

Note : le niveau global des indices est fixé en référence à l'agglomération parisienne (recalculé par les auteurs à partir des données publiées pour référence à l'agglomération parisienne).

caractéristiques fortement structurelles qui évoluent donc assez peu au cours du temps ; les prix sont plus élevés en Corse, probablement en raison de la géographie insulaire qui limite la concurrence et renchérit les coûts de production notamment par les coûts de transport des produits élaborés sur le continent ; ils sont aussi plus élevés en région parisienne, cette fois sans doute en raison de coûts de commercialisation plus élevés (prix de l'immobilier commercial) et du pouvoir d'achat des résidents-consommateurs, en moyenne plus élevé qu'ailleurs.

### Les données

Les données utilisées sont les données de caisse des enseignes signataires d'une convention avec l'Insee autorisant l'institut à accéder à des enregistrements quotidiens pour l'année 2013. Dans ces données, on ne retient pour cette étude que celles qui se rapportent à l'alimentaire industriel, c'est-à-dire aux produits d'alimentation et aux boissons, alcoolisées ou non<sup>7</sup>, vendus en supermarchés. Ces données sont issues, en avril 2013, de 1 833 magasins. Ces magasins sont répartis dans 1 330 communes situées dans 707 agglomérations de métropole. La répartition du nombre

de points de vente par grandes agglomérations<sup>8</sup> dans les études relatives plus haut est donnée au tableau 2.

La répartition par région est donnée dans le tableau 3. Notons qu'il s'agit, ici comme dans tout l'article, des régions administratives antérieures à la réforme (loi NOTRe) de 2015. Dans l'ensemble, la distribution du nombre de points de vente au niveau régional est relativement proche de celle de la démographie. Ainsi les points de vente de la base de données renvoient, par leur répartition géographique, une image assez fidèle de l'espace marchand français. Naturellement, dans la mesure où un nombre limité de groupes de la grande distribution ont transmis leurs données à l'Insee en 2013, des effets de grappes restent à craindre.

La structure de consommation, en termes de produits consommés, devrait *a priori* être voisine d'une région à l'autre. Afin d'examiner cette hypothèse,

7. Division de la nomenclature COICOP 01, hors produits frais (fruits et légumes frais, crustacés ainsi qu'une partie des poissons et des viandes) et groupe de la nomenclature COICOP 02.1.

8. Nomenclature des unités urbaines, version 2010. Cette nomenclature comprend environ 2 000 unités.

**Tableau 2**  
**Nombre de points de ventes par grande agglomération dans l'échantillon utilisé**

Agglomération	Nombre de points de vente
Agglomération parisienne	352
Lyon	50
Marseille	31
Bordeaux	30
Rennes	10
Reims	8
Rouen	15
Strasbourg	19
Lille	26
Orléans	13
Limoges	4
Ajaccio-Bastia	4
Clermont-Ferrand	16
Toulouse	26
Dijon	4
Nantes	9
Nancy	5
Poitiers	2
Montpellier	12

Note : lorsque le nombre de points de vente est inférieur ou égal à 4 (Limoges, Dijon, Poitiers, Ajaccio-Bastia), l'indice de la ville ne figure pas dans la table de résultats (voir tableau 7).

Source : Insee, données de caisse 2013.

nous procédons au calcul de cette structure avec la base des données de caisse : le tableau 4 donne la répartition des chiffres d'affaires associés aux regroupements des produits selon la nomenclature internationale par fonctions de consommation (COICOP). Ces statistiques montrent, comme attendu, que les structures régionales, dans le champ de l'alimentaire industriel, s'écartent peu d'une structure moyenne métropolitaine portant sur le même champ. On note également que cette structure, propre aux achats réalisés en supermarchés, diffère sensiblement de la structure de consommation toutes formes de ventes confondues, essentiellement s'agissant des produits frais non industriels (fruits et légumes frais, crustacés, une partie des poissons et des viandes).

Ainsi construite, la base de données comprend, en moyenne, 16.4 millions d'observations par semaine, correspondant au croisement [point de vente × EAN], des prix moyens par codes-barres et des chiffres d'affaires. Le chiffre d'affaires total d'une semaine d'observation disponible dans la base de données est, en moyenne, d'environ 445 millions d'euros. Extrapolé sur une année (52 semaines) et rapporté à la

**Tableau 3**  
**Nombre de points de ventes par région dans l'échantillon utilisé**

Région	Nombre de points de vente	Poids en nombre de PV (en %)	Poids démographique (en %)
Île-de-France	404	22.1	18.8
Rhône-Alpes	201	11.0	10.0
Nord-Pas-de-Calais	162	8.9	6.4
Provence-Alpes-Côte d'Azur	105	5.7	7.8
Centre	104	5.7	4.0
Aquitaine	94	5.1	5.2
Haute-Normandie	79	4.3	2.9
Picardie	73	4.0	3.0
Midi-Pyrénées	72	3.9	4.6
Bretagne	71	3.9	5.1
Auvergne	67	3.7	2.1
Languedoc-Roussillon	65	3.6	4.2
Basse-Normandie	58	3.2	2.3
Pays de la Loire	51	2.8	5.7
Lorraine	44	2.4	3.7
Alsace	44	2.4	2.9
Champagne-Ardenne	36	2.0	2.1
Bourgogne	33	1.8	2.6
Limousin	25	1.4	1.2
Poitou-Charentes	21	1.1	2.8
Franche-Comté	15	0.8	1.9
Corse	5	0.3	0.5
Total	1 829	100	100

Lecture : dans la base de données utilisée, l'Île-de-France comprend 404 points de vente. Les 404 points de vente représentent 22.1 % des 1 829 points de vente que comprend la base de données. Pour rappel et comparaison, l'Île-de-France regroupe 18.8 % des habitants de France métropolitaine (Recensement de la population, 2012). Les chiffres donnés en italiques ne proviennent pas de la base de données de caisse.

Source : Insee, données de caisse 2013.

dépense<sup>9</sup> de consommation des ménages observée en 2012 dans l'alimentaire et les boissons alcoolisées ou non, ce chiffre d'affaires représente environ 15 % de la dépense de consommation des ménages du champ<sup>10</sup>.

### Modèle d'estimation

Une observation élémentaire correspond à un code-barres (EAN) vendu dans un magasin de l'échantillon durant la semaine considérée. Il y a donc une observation par [point de vente × EAN]. On suppose que les observations élémentaires ainsi définies sont repérées par un indice  $i$  d'un ensemble  $I$ . Ainsi  $p_i$  est le prix (valeur unitaire sur la semaine) de l'article repéré par son code-barres dans un des magasins de la base de données. On note  $\omega_i$  le chiffre d'affaires associé à l'observation correspondante.

L'indice traduisant les écarts de niveaux de prix entre zones géographiques est calculé par

une méthode hédonique (Triplett, 2006). Cette approche, fondée sur une modélisation économétrique des prix, diffère quelque peu des approches harmonisées appliquées pour la mesure des parités de pouvoir d'achat entre pays européens. Pour autant, elle figure au nombre des méthodes classiques (Deaton & Heston, 2010) et, dans le cas où les territoires comparés présentent des consommations proches (en prix et en structure – c'est le cas ici comme le montre le tableau 4), elle conduit à des mesures d'écarts de niveaux de prix voisines des méthodes alternatives.

Le modèle économétrique est conditionné par le code-barres et la zone géographique d'appartenance du produit  $i$  considéré. Grâce au conditionnement par le code-barres, le modèle

9. Source Comptabilité nationale, euros courants, soit 156 milliards d'euros.  
10. Pour être précis, les écarts sur le champ tiennent aux produits alimentaires vendus dans les autres points de vente (de la Grande distribution pour des enseignes non intégrées dans l'étude car ne transmettant pas en 2013 leurs données à l'Insee, ainsi que d'autres types de magasins, ou des marchés) et aux produits frais.

Tableau 4  
Structures régionales de consommation dans le champ de l'alimentaire industriel

Région	Code	01.1.1	01.1.2	01.1.3	01.1.4	01.1.5	01.1.6	01.1.7	01.1.8	01.1.9	01.2.1	01.2.2	02.1.1	02.1.2	02.1.3	Total
Île-de-France	11	13.3	10.2	5.4	19.4	3.0	1.1	6.0	7.7	2.6	3.4	11.0	5.4	9.1	2.6	100
Champagne-Ardenne	21	9.8	10.4	4.1	17.2	2.8	0.9	5.6	6.1	2.0	3.2	9.7	6.3	18.0	3.8	100
Picardie	22	10.7	11.6	4.6	18.3	3.4	0.8	5.9	6.1	2.2	3.3	11.0	9.1	9.2	3.7	100
Haute-Normandie	23	10.6	10.7	4.5	17.0	3.1	0.9	5.7	6.4	2.1	3.5	10.3	11.4	10.6	3.2	100
Centre	24	11.3	11.1	5.2	18.8	3.4	1.0	6.1	6.8	2.2	3.5	10.5	8.0	8.3	3.7	100
Basse-Normandie	25	11.1	9.7	4.5	17.5	3.3	1.0	5.8	7.0	2.0	3.8	9.0	9.5	12.6	3.3	100
Bourgogne	26	10.7	10.7	4.6	18.5	3.2	0.9	6.0	6.9	2.3	3.5	10.1	6.5	12.3	3.8	100
Nord-Pas-de-Calais	31	9.8	10.0	4.0	16.9	3.4	0.8	5.4	6.4	2.2	3.3	11.9	8.4	12.7	4.6	100
Lorraine	41	11.6	10.2	4.7	19.9	3.2	0.8	5.8	7.0	2.4	3.8	12.0	4.8	9.0	4.8	100
Alsace	42	11.7	9.3	4.6	19.8	3.5	1.0	5.6	7.2	2.9	3.8	13.5	4.4	7.6	5.2	100
Franche-Comté	43	11.1	10.3	5.1	17.9	3.3	1.0	6.1	7.2	2.3	3.9	10.3	5.5	11.6	4.6	100
Pays de la Loire	52	11.9	10.2	5.0	17.9	3.4	1.1	6.2	7.2	2.1	3.5	9.5	7.8	10.1	4.2	100
Bretagne	53	11.3	10.4	4.2	16.5	3.4	1.2	5.7	7.3	2.0	3.7	8.9	7.2	14.2	4.0	100
Poitou-Charentes	54	10.6	11.3	5.3	18.2	3.2	1.0	5.9	6.4	2.1	3.6	10.2	7.2	10.7	4.2	100
Aquitaine	72	11.6	10.6	5.7	18.7	3.3	1.1	6.4	7.0	2.2	4.0	10.1	5.5	9.5	4.3	100
Midi-Pyrénées	73	12.5	9.8	5.7	19.3	3.3	1.1	6.2	7.6	2.5	4.1	10.0	5.1	8.7	4.4	100
Limousin	74	10.5	9.7	4.8	17.7	3.4	1.1	5.7	6.9	2.1	3.8	9.4	7.5	13.3	4.2	100
Rhône-Alpes	82	12.4	9.7	5.4	18.9	3.3	1.0	5.7	7.8	2.6	3.5	10.3	5.3	9.9	4.1	100
Auvergne	83	11.8	10.1	5.0	17.8	3.7	1.0	5.9	7.9	2.3	4.0	9.8	7.1	9.4	4.4	100
Languedoc-Roussillon	91	12.1	10.9	5.7	19.9	3.2	1.0	6.1	7.3	2.6	4.3	10.4	4.6	8.1	3.9	100
Provence-Alpes-Côte d'Azur	93	11.7	10.4	5.9	19.8	3.2	1.0	5.7	6.9	2.6	3.8	10.1	5.4	10.3	3.4	100
Corse	94	12.6	11.9	6.7	19.4	3.4	1.1	7.3	7.4	2.7	4.1	8.2	4.5	8.1	2.7	100
France métropolitaine (1)		11.9	10.3	5.1	18.7	3.2	1.0	5.9	7.2	2.4	3.6	10.6	6.4	10.2	3.6	100
France (2)		14.3	21.6	5.2	12.1	1.8	5.8	9.8	6.8	3.4	2.2	5.4	4.1	5.7	1.7	100

Note : répartition (%) territoriale du chiffre d'affaires, selon le type de produit, par regroupement de classes de la nomenclature COICOP. 01.1.1 : Pain et céréales ; 01.1.2 : Viande ; 01.1.3 : Poissons et crustacés ; 01.1.4 : Lait, fromage et œufs ; 01.1.5 : Huiles et graisses ; 01.1.6 : Fruits ; 01.1.7 Légumes ; 01.1.8 : Sucre, confitures, chocolat, confiserie et produits glacés ; 01.1.9 : Sel, épices, sauces et produits alimentaires non ailleurs ; 01.2.1 : Café, thé et cacao ; 01.2.2 : Autres boissons non alcoolisées ; 02.1.1 : Alcools ; 02.1.2 : Vins, cidres et champagne ; 02.1.3 : Bières. Calcul des auteurs sur la base des données de caisse de la semaine de référence (avril 2013), y compris pour (1). (2) répartition France entière, Comptabilité nationale (tableaux détaillés de consommation des ménages pour 2013).  
Source : Insee, données de caisse 2013.

retenu permet d'estimer les écarts moyens de prix entre zones géographiques. Formellement, on postule que le prix  $p_i$  répond à un processus générateur de la forme :

$$\log(p_i) = c + \sum_{\ell=1}^L \alpha_{\ell} \cdot \mathbf{1}_{(ean_i=\ell)} + \sum_{z=1}^Z \beta_z \cdot \mathbf{1}_{(zone_i=z)} + \varepsilon_i \quad (1)$$

où  $\mathbf{1}$  désigne une variable indicatrice valant 1 si la condition figurant entre parenthèses en indice est vraie et 0 sinon,  $ean_i$  est le numéro de code-barres de l'observation  $i$  et  $zone_i$  est la zone géographique à laquelle appartient l'observation  $i$ .  $\varepsilon_i$  est un aléa centré. Dans ce modèle, les coefficients  $c$ ,  $\alpha_{\ell}$  ( $\ell \in \{1, \dots, L\}$ ,  $L$  est le nombre de codes-barres pris en compte) et  $\beta_z$  ( $z \in \{1, \dots, Z\}$ ,  $Z$  est le nombre de zones géographiques prises en compte) sont inconnus. On les estime par moindres carrés. Les ratios<sup>11</sup> des coefficients  $\alpha_{\ell}$  s'interprètent comme les rapports de prix moyens associés aux codes-barres considérés. Les rapports des coefficients  $\beta_z$  traduisent les rapports de prix moyens entre zones géographiques, à produits donnés (repérés par leurs codes-barres). Ces coefficients,

estimés par moindres carrés, correspondent à des indices de prix hédoniques (Triplett, 2006 ; Diewert, 2003 ; Silver & Heravi, 2005).

La forme des estimateurs obtenus est détaillée dans l'encadré. On constate que l'estimateur obtenu fait naturellement intervenir les différences de structure de consommation entre régions, par l'intermédiaire des pondérations utilisées. La pondération la plus naturelle, de ce point de vue, est par le chiffre d'affaires du produit dans le point de vente considéré. Le modèle de référence fait donc intervenir une pondération par le chiffre d'affaires. La pondération unitaire fait *de facto* intervenir une structure, assez proche de celle par le chiffre d'affaires, puisqu'elle se fonde sur le nombre de transactions pour le produit et le point de vente considérés. L'approche alternative par la pondération unitaire est donc utilisée afin d'examiner la robustesse des résultats vis-à-vis de la pondération de référence.

11. Le rapport des exponentielles de ces coefficients pour être précis (voir infra).

#### ENCADRÉ – Structure des estimateurs hédoniques

L'estimateur des moindres carrés (1) peut ou non être pondéré. Concrètement, il y a deux options possibles : soit on utilise des pondérations s'apparentant aux chiffres d'affaires  $\omega_i$ , soit on ne pondère pas les observations élémentaires. Afin de bien évaluer les conséquences du choix que nous faisons en termes de pondérations, il est utile d'examiner la forme des estimateurs que nous obtenons pour les coefficients  $\beta_z$ . Pour cela, nous supposons pour plus de simplicité que l'estimation est réalisée en deux étapes<sup>(a)</sup> : une première étape dans laquelle les coefficients  $\alpha_{\ell}$  sont estimés. Puis, dans une seconde étape, les coefficients  $\beta_z$  sont estimés (conditionnellement aux estimateurs  $\hat{\alpha}_{\ell}$  des  $\alpha_{\ell}$  obtenus en première étape). Évidemment, en procédant de la sorte, nous n'obtenons pas l'estimateur des moindres carrés que nous obtiendrions si les vecteurs  $(\alpha, \beta)$  étaient estimés simultanément, mais les limites en probabilité des deux estimateurs en deux étapes sont les mêmes que celles de l'estimateur en une étape<sup>(b)</sup>. L'intérêt de procéder en deux étapes est que l'on peut aisément dériver la forme de  $\hat{\beta}$ . En effet, soit  $\tilde{p}_i$  la variable  $p_i$  corrigée de la première étape :

$$\log(\tilde{p}_i) = \log(p_i) - c - \sum_{\ell=1}^L \hat{\alpha}_{\ell} \cdot \mathbf{1}_{(ean_i=\ell)} \quad (2)$$

alors la deuxième étape consiste à régresser  $\log(\tilde{p}_i)$  sur les vecteurs ligne  $x_i$  comprenant  $Z$  colonnes, dont  $Z-1$  sont nulles, et la seule non nulle est égale à 1 :

$$\log(\tilde{p}_i) = x_i \cdot \beta + v_i \quad (3)$$

L'estimateur des moindres carrés  $\hat{\beta}$  est classiquement solution de l'équation (ici en version pondérée ; pour une version non pondérée, il suffit de poser  $\omega_i = 1$ ) :

$$\left( \sum_{i \in I} \omega_i x_i' \cdot x_i \right) \cdot \hat{\beta} = \sum_{i \in I} \omega_i x_i' \cdot \log(\tilde{p}_i)$$

soit, en regroupant par modalité de zone<sup>(c)</sup> :

$$\text{Diag} \begin{pmatrix} \sum_{i \in z_1} \omega_i \\ \vdots \\ \sum_{i \in z_z} \omega_i \end{pmatrix} \cdot \hat{\beta} = \begin{pmatrix} \sum_{i \in z_1} \omega_i \log(\tilde{p}_i) \\ \vdots \\ \sum_{i \in z_z} \omega_i \log(\tilde{p}_i) \end{pmatrix}$$

et finalement, pour la zone  $k$  considérée (aussi notée  $z_k$ ) :

$$\exp(\hat{\beta}_k) = \left\{ \prod_{i \in z_k} \tilde{p}_i^{\omega_i} \right\}^{1/\sum_{i \in z_k} \omega_i} \quad (4)$$

Il en découle que, pour les zones  $k$  et  $j$ , nous avons :

$$\exp(\hat{\beta}_j - \hat{\beta}_k) = \frac{\left\{ \prod_{i \in z_j} \tilde{p}_i^{\omega_i} \right\}^{1/\sum_{i \in z_j} \omega_i}}{\left\{ \prod_{i \in z_k} \tilde{p}_i^{\omega_i} \right\}^{1/\sum_{i \in z_k} \omega_i}} \quad (5)$$

On note que ce ratio correspond à un rapport de prix moyens<sup>(d)</sup> (i.e. ratio de valeurs unitaires). On observe que cet indice d'écart de niveau de prix entre zones prend en compte les structures locales de consommation puisque, tant au numérateur qu'au dénominateur, chaque produit pèse dans l'indice en proportion de son poids dans la dépense de consommation locale.

(a) Cette décomposition en deux étapes est uniquement proposée ici dans le but d'explicitier la forme de l'indice obtenu. En pratique, nous faisons un calcul en une seule étape fondé sur le modèle (1).

(b) Sous les mêmes hypothèses de convergence, notamment d'orthogonalité de l'aléa et des variables explicatives.

(c) *Diag* désigne la matrice diagonale dont la diagonale coïncide avec le vecteur en argument.

(d) En moyenne géométrique, à interpréter comme étant calculé à code-barres fixé, identique pour le numérateur et le dénominateur, en raison du conditionnement par l'EAN dans les étapes (1) et (2).

À ce stade, il convient de préciser les conditions sous lesquelles l'estimateur  $\hat{\beta}_k$  est non biaisé. En tant qu'estimateur du coefficient  $\beta_k$  de l'équation (1), le coefficient est non biaisé dès que les conditions d'orthogonalité des variables explicatives et de l'aléa  $\varepsilon_i$  (ou  $v_i$  dans le cas de la régression de deuxième étape) sont assurées. On fait l'hypothèse ici que c'est bien le cas. En revanche, la statistique  $\exp(\hat{\beta}_k)$  n'est pas un estimateur sans biais de  $\exp(\beta_k)$ . En effet, à partir de l'expression de l'estimateur des moindres carrés (équation 1 ou 3), on montre<sup>12</sup> que :

$$E \left[ \frac{p_i}{p_j} \middle| i \in z_k, \ell \in z_j, ean_i = ean_j \right] = \exp(\beta_k - \beta_j) [1 + \sigma^2] \quad (6)$$

où  $\sigma^2$  est la variance des  $\varepsilon_i$  qu'on supposera désormais de variance identique. Nous utiliserons donc cette correction pour calculer les rapports de prix.

## Résultats

### Les écarts observés en avril 2013

Dans cette partie, nous présentons les résultats fondés sur des régressions du type du modèle (1) pour une semaine de données en avril 2013 (troisième semaine du mois). D'un point de vue pratique, pour toutes les régressions réalisées, on ne conserve que 5 000 références de codes-barres par enseigne. On choisit, parmi les références vendues de l'enseigne, les 5 000 principales en termes de chiffre d'affaires. En effet, le modèle

hédonique (1) est fondé sur des indicatrices de codes-barres. Celles-ci ne sont pas explicitement estimées (elles sont réduites algébriquement dans l'équation normale), mais un trop grand nombre de références conduit à une équation normale trop lourde à traiter. Différents tests ont été menés pour examiner les conséquences de cette restriction. L'expérience montre que retenir 3 000 ou 5 000 références par enseigne ne conduit pas, sur les indicatrices géographiques, à des résultats notablement différents. Au final, la réunion, pour l'ensemble des enseignes de la base, des 5 000 principaux codes-barres les concernant, conduit à considérer 13 098 codes-barres dans les régressions. Ce nombre étant nettement plus élevé que les 5 000 références conservées par enseigne, une fraction significative des codes-barres est donc spécifique<sup>13</sup> aux enseignes (marques de distributeur). Compte-tenu de cette restriction, la base de calcul comprend 7.3 millions d'enregistrements correspondant aux croisements [point de vente  $\times$  codes-barres retenus]. En termes de chiffre d'affaires, la restriction effectuée conduit à conserver 74 % de l'information contenue dans la base d'origine présentée dans la section sur les données.

Le tableau 5 indique, par type de produits et pour la base de données restreinte aux 5 000 principaux

12. Par exemple par une  $\Delta$ -méthode ou bien en faisant des hypothèses sur la distribution normale des aléas dans l'équation (1).  $E$  désigne l'espérance mathématique (notation conditionnelle).

13. Si chaque code-barres était vendu dans toutes les enseignes, la réunion des 5 000 principaux codes-barres d'enseignes comprendrait précisément 5 000 codes-barres.

Tableau 5  
Répartition des familles IRI et des codes-barres par regroupement de la nomenclature COICOP

Code COICOP	Libellé COICOP	Nombre de familles	Nombre de codes-barres
0111	Pain	47	2 200
0112	Viande	19	1 479
0113	Poissons et crustacés	22	848
0114	Lait, fromage et œufs	23	1 830
0115	Huiles et graisses	6	300
0116	Fruits	15	252
0117	Légumes	31	1 117
0118	Sucre, confitures, chocolat, confiseries et produits glacés	29	1 098
0119	Sel, épices sauces et autres produits alimentaire	35	564
0121	Café, thé et cacao	10	409
0122	Autres boissons non alcoolisées	17	876
0211	Alcools	12	361
0212	Vins, cidres et champagne	21	1 535
0213	Bières	1	229
Total		288	13 098

Lecture : la base de données comprend 47 familles de produits IRI appartenant au regroupement COICOP 0111 (Pains). 2 200 références codes-barres s'y rapportent dans la base de données examinée.

Source : Insee, données de caisse 2013.

codes-barres par enseigne, le nombre de famille de produits IRI<sup>14</sup> qui s'y rattachent, ainsi que le nombre de références code-barres correspondant. Grossièrement, une famille IRI correspond à un type de produit dont le grain est approximativement de même finesse que celui des variétés de produits suivies dans l'IPC (Insee, 1998). Pour mémoire, 327 variétés sont suivies dans l'IPC métropolitain au titre de l'alimentaire industriel en 2013. Ce chiffre est effectivement comparable au nombre de familles IRI qui, sur le même champ, est de 288. Dans la base de données étudiée, le nombre de références code-barres correspondant est, comme vu plus haut, de 13 098.

Le tableau 6 donne les résultats d'estimation des indices d'écart de niveaux de prix dans l'alimentaire industriel pour les régions administratives métropolitaines, calculés à l'aide de la base de données de caisse. On constate, tout d'abord, que les écarts sont relativement peu dispersés : 5.5 à 8 points de pourcentage selon que l'on pondère ou non les observations par leurs chiffres d'affaires. La dispersion est plus importante lorsqu'on considère les indices non pondérés plutôt que les

indices pondérés. Cela suggère que les produits pesant davantage dans le budget des consommateurs connaissent une dispersion spatiale de prix plus faible que les autres produits. Il est aussi remarquable de constater que l'ordre des régions classées par niveau d'écart moyen de prix n'est pas modifié selon que l'on pondère ou non les observations par le chiffre d'affaires.

Sur un plan géographique, les résultats dessinent des dominantes régionales : un grand Centre-Ouest de la France dans lequel les niveaux de prix sont environ 3 % plus bas qu'en Île-de-France ; puis une partie comprenant les régions plus rurales du Centre, celles du Nord de la France ainsi que l'Aquitaine dans lesquelles les prix de l'alimentaire industriel sont en moyenne de 2 % inférieurs à ceux de l'Île-de-France ; les régions plus industrielles et urbaines de l'Est et du Sud présentent des niveaux de prix alimentaires 1 % plus bas

14. Compagnie privée qui développe un catalogue, utilisé par l'Insee dans le cadre de l'expérience pilote, de caractéristiques de produits référencés par codes-barres.

Tableau 6  
Indices d'écart de niveau de prix entre la région parisienne et les autres régions

Région	Code	Estimation		
		Pondérée	Non pondérée	Pondérée avec E.F. d'enseigne
Bretagne	53	96.7	95.4	97.1
Pays de la Loire	52	97.0	96.1	97.6
Centre	24	97.6	96.8	97.9
Limousin	74	97.8	96.5	98.0
Poitou-Charentes	54	97.4	96.6	98.2
Basse-Normandie	25	97.9	96.8	98.2
Auvergne	83	98.2	97.2	98.4
Haute-Normandie	23	98.1	97.5	98.4
Midi-Pyrénées	73	98.3	97.2	98.4
Nord-Pas-de-Calais	31	97.9	97.1	98.6
Bourgogne	26	97.7	96.9	98.6
Picardie	22	98.2	97.4	98.6
Aquitaine	72	98.2	97.3	98.6
Franche-Comté	43	97.9	97.1	98.7
Champagne-Ardenne	21	98.1	97.4	98.7
Alsace	42	98.9	98.5	98.9
Lorraine	41	98.6	98.0	99.0
Languedoc-Roussillon	91	98.6	98.0	99.2
Rhône-Alpes	82	98.9	98.2	99.3
Provence-Alpes-Côte d'Azur	93	99.2	98.9	99.9
Île-de-France	11	100 (Réf.)	100 (Réf.)	100 (Réf.)
Corse	94	102.1	103.5	102.8

Lecture : au sens de l'estimation dans laquelle les observations sont pondérées par leur chiffre d'affaires, les prix sont en moyenne 3.3 % plus bas en Bretagne qu'en Île-de-France. Au sens de l'estimation dans laquelle les observations sont pondérées unitairement, les prix sont en moyenne 4.4 % plus bas en Bretagne qu'en Île-de-France. Les indicatrices de zones résultent d'une régression de type (1) dans laquelle les zones sont les anciennes régions administratives. La dernière colonne se réfère à un calcul équivalent à celui réalisé pour la première colonne (i. e. pondéré), dans lequel un effet-fixe enseigne a été ajouté. Les résultats obtenus sont corrigés conformément à la formule (6) et transformés en indices par une multiplication par 100. La variance estimée de l'aléa est de 0.004. Calcul effectué sur 7.3 millions d'enregistrements. L'écart-type moyen sur les indices présentés est de 0.02 point d'indice. Source : Insee, données de caisse 2013.

qu'en Île-de-France. Enfin, les prix en Corse sont 2 % plus élevés qu'en Île-de-France.

Afin de comparer aux résultats « historiques » présentés au tableau 1, le tableau 7 regroupe les indices d'écarts de prix alimentaires industriels entre les grandes agglomérations de métropole et l'agglomération parisienne. Pour comparer ces résultats à ceux du tableau 1, il convient de rappeler que les champs économiques et géographiques ainsi que les méthodes de calcul ne sont pas rigoureusement homogènes. Une partie des écarts constatés entre agglomérations et leurs évolutions au cours du temps intègrent vraisemblablement des biais liés à l'inhomogénéité de champ et de méthode. Néanmoins, il est quand même intéressant d'examiner les résultats obtenus.

Pour les agglomérations comme pour les régions, on constate (cf. tableaux 6 et 7) que les écarts de niveaux de prix estimés par régression non pondérée sont un peu plus importants que ceux

calculés par régression pondérée. Hors Corse<sup>15</sup>, les écarts de prix se situent dans une fourchette de 3.7 à 4.4 points de pourcentage selon qu'on pondère ou non les observations. Par rapport à l'agglomération parisienne dans laquelle les prix sont les plus élevés, les agglomérations les moins chères (parmi les grandes agglomérations) pour l'alimentaire industriel sont Nantes, Rennes, Orléans, Rouen et Lille. Remarquablement, c'était déjà le cas en 1989 (Insee, Division prix de détail 1990) et en 1985 (Mineau, 1987) – cf. tableau 1. La distance avec le schéma de 1977 (Baraille, 1978) est légèrement plus marquée.

Par rapport aux écarts mis en évidence entre régions, ceux constatés entre grandes agglomérations sont un peu plus accentués. Par exemple, en référence à une zone quasi-comparable

15. Non présentée dans le tableau des agglomérations (tableau 7) en raison d'un nombre de points de vente trop restreint dans la base de données de caisse.

**Tableau 7**  
**Écarts de niveau de prix entre l'agglomération de Paris et les principales autres agglomérations métropolitaines**

Agglomération	Estimation	
	Pondérée	Non pondérée
Agglomération parisienne	100 (Réf.)	100 (Réf.)
Lyon	98.6	97.7
Marseille	98.9	98.4
Bordeaux	97.9	97.0
Rennes	96.5	95.6
Reims	97.9	97.6
Rouen	97.1	96.6
Strasbourg	99.1	98.7
Lille	97.3	96.5
Orléans	97.1	95.6
Limoges	nd	nd
Ajaccio-Bastia	nd	nd
Clermont-Ferrand	98.4	97.3
Toulouse	98.0	96.7
Dijon	nd	nd
Nantes	96.3	95.9
Nancy	98.4	97.7
Poitiers	nd	nd
Montpellier	97.9	97.1
Limoges	96.6	95.8
Ajaccio-Bastia	101.5	102.3
Dijon	97.1	96.5
Poitiers	97.7	97.0

Lecture : au sens de l'estimation dans laquelle les observations sont pondérées par leur chiffre d'affaires, les prix sont en moyenne 1.4 % plus bas à Lyon qu'à Paris. Au sens de l'estimation dans laquelle les observations sont pondérées unitairement, les prix sont en moyenne 2.3 % plus bas à Lyon qu'à Paris. Les indicatrices de zones résultent d'une régression de type (1) dans laquelle les zones sont les agglomérations (unités urbaines). Les résultats obtenus sont corrigés conformément à la formule (6) et transformés en indices par une multiplication par 100. La variance estimée de l'aléa est de 0.004. Calcul effectué sur 7.3 millions d'enregistrements. L'écart-type moyen sur les indices présentés est de 0.10 point d'indice. Source : Insee, données de caisse 2013.

(l'agglomération parisienne ou l'Île-de-France selon le cas), l'indice (pondéré) de Montpellier est de 97.9 tandis que celui du Languedoc-Roussillon est de 98.6. De même, celui de Lille est de 97.3 tandis que celui du Nord-Pas-de-Calais est de 97.9. Cette situation<sup>16</sup> pourrait être liée au fait que la concurrence est probablement plus forte au sein des marchés locaux de grandes agglomérations, ce qui aurait tendance à tirer les prix vers le bas.

Cette règle souffre toutefois de deux exceptions parmi les grandes agglomérations : il s'agit de Strasbourg dont l'indice est de 99.1 tandis que celui de l'Alsace est de 98.9 et Clermont-Ferrand dont l'indice est de 98.4 tandis que celui de l'Auvergne est de 98.2. Dans ces deux cas, les écarts ne sont néanmoins pas significatifs.

Comme évoqué en introduction et plus haut, la représentativité de l'échantillon de données par rapport à la distribution spatiale des prix peut être altérée du fait du nombre limité d'enseignes ayant fourni leurs données à l'Insee en 2013. Ainsi, il est possible que la sélection des enseignes de l'échantillon soit corrélée à la dimension régionale sur laquelle sont estimées les statistiques proposées. C'est le cas, par exemple, si une enseigne de l'échantillon dont la politique de prix diffère des autres (disons par exemple que ses prix sont systématiquement moins élevés) est, du fait de la sélection, surreprésentée dans une région et pas ailleurs. Dans ce cas, l'estimation du niveau des prix dans la région concernée par la surreprésentation est biaisée (à la baisse pour l'exemple) par rapport aux autres régions.

Pour démontrer l'absence ou l'existence d'un tel biais et pour l'évaluer, il faudrait disposer d'un jeu complet de données relatif à l'ensemble des enseignes. Avec l'échantillon limité dont on dispose, s'il n'est pas possible de procéder à une étude définitivement concluante sur ce point, il est toutefois possible d'examiner la cohérence de certains résultats avec l'hypothèse de représentativité du sous-échantillon étudié. Le premier résultat utile à cet égard a déjà été présenté dans le tableau 3 qui montre que la distribution régionale des points de vente respecte la distribution de la population et donc, vraisemblablement, la dépense de consommation des ménages dans l'alimentaire. Un autre résultat utile montre l'intérêt d'ajouter dans l'équation (1) des indicatrices d'enseignes. Si, par exemple, un indice régional est significativement modifié dans ce second calcul par rapport au calcul de référence, c'est le signe que le niveau régional des prix tient pour partie aux enseignes représentées dans le réseau local des supermarchés, dans le sous-échantillon étudié. Il

est possible, dans ces conditions, que la portée des résultats obtenus soit, au fond, limitée au seul échantillon considéré. Nous avons procédé à un tel calcul dont les résultats, en termes d'indices régionaux pondérés par les ventes, sont présentés au tableau 6 en dernière colonne. Ces résultats sont à comparer avec ceux du calcul de référence (en gras dans ce même tableau). Il s'avère que les indices régionaux peuvent être assez nettement modifiés, jusqu'à 0.8 points pour la Bourgogne et la Franche-Comté. Ceci étant, les constats principaux, en particulier la hiérarchie des prix entre la Corse, l'Île-de-France et les autres régions métropolitaines, de même que l'ordre de grandeur des écarts, demeurent.

Finalement, si des « effets enseignes » existent, leur impact dans les indices locaux étant perceptible, la généralisation, à l'ensemble de la consommation alimentaire dans la grande distribution, des principaux enseignements tirés du sous-échantillon étudié est raisonnablement accréditée par les différents tests de robustesse réalisés.

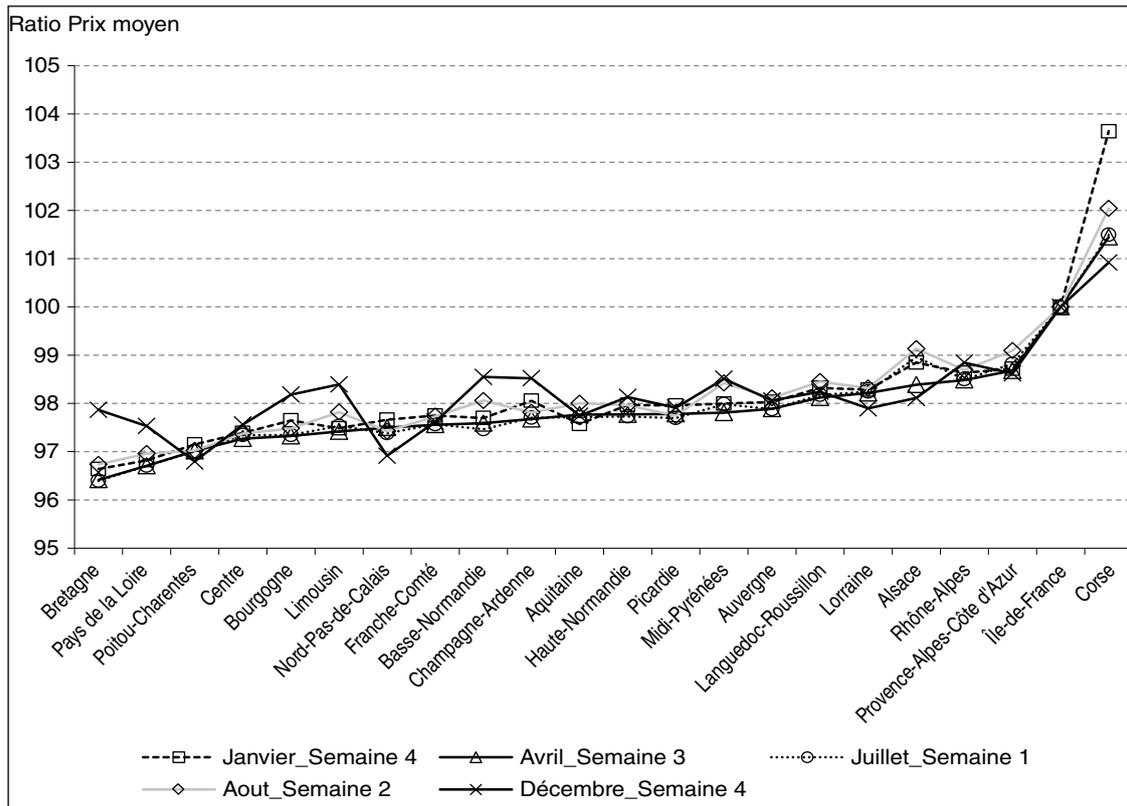
### **Sensibilité des résultats au choix de la semaine de calcul**

Afin de tester la robustesse des résultats obtenus, nous examinons maintenant le comportement des écarts régionaux de niveaux de prix lorsque la semaine retenue change. Pour cela, nous généralisons l'analyse précédente à quatre autres semaines de l'année 2013 relativement typiques quant à leur situation par rapport aux soldes et aux fêtes à fort impact sur les achats de consommation : quatrième semaine de janvier (peu après les fêtes de fin d'année et en pleine période de soldes d'hiver), première semaine de juillet (début des vacances d'été), deuxième semaine d'août (fin des vacances d'été) et quatrième semaine de décembre (fêtes de fin d'année). Ces semaines sont comparées à la troisième semaine d'avril, précédemment étudiée et servant de référence à la comparaison.

La figure suivante montre les écarts de niveaux de prix entre l'Île-de-France et les autres régions pour les 5 semaines étudiées, les régions étant ordonnées en abscisse selon leur rang en termes de niveau de prix observé lors de la semaine de référence d'avril. On constate que les écarts sont

16. Pour l'interprétation, on fait ici l'hypothèse, raisonnable compte tenu de son poids, que le niveau de prix de l'agglomération parisienne est aussi celui de l'Île-de-France. Par conséquent, les écarts d'indices des villes de province et de leur région sont liés à des différences locales entre les villes et leur région d'appartenance et non à d'éventuelles différences de prix entre l'unité urbaine de Paris et sa région.

Figure  
Écart de niveaux de prix régionaux par rapport à l'Île-de-France



Note : référence 100 pour l'Île-de-France pour chacune des semaines d'étude. L'ordre des régions figurant en abscisses est croissant selon le niveau d'indice constaté en avril 2013.

Source : Insee, données de caisse 2013.

extrêmement voisins d'une semaine à l'autre. Il existe toutefois deux exceptions à cette proximité. D'une part, le niveau des prix en Corse est relativement plus élevé en janvier que lors des autres semaines examinées. D'autre part, on note une structure de prix régionale assez nettement modifiée lors de la dernière semaine de décembre ce qui est l'effet probable de la spécificité des produits vendus à cette époque et des mouvements de population importants lors des fêtes qui modifient la structure géographique des marchés.

Au final, cette analyse de robustesse tend à confirmer le caractère essentiellement structurel des écarts géographiques de niveaux de prix. Elle montre également la richesse des lots de données de caisse qui permettent d'estimer avec une grande précision des indices de prix sur des pas géographiques ou temporels inaccessibles aux moyens d'enquête classiques.

\* \*  
\*

Cette étude présente un exemple d'exploitation des données de caisse pour établir des mesures de différences de niveaux de prix entre territoires métropolitains dans l'alimentaire et les boissons alcoolisées ou non. Naturellement, compte-tenu du champ couvert par les données de caisse utilisées, la portée de ces résultats reste limitée et leur généralisation à toute la consommation alimentaire des ménages métropolitains est sujette à discussion. D'abord parce que les enseignes concourant à l'expérience pilote menée par l'Insee en 2013 sont relativement peu nombreuses (même si elles contribuent à hauteur de 30 % du chiffre d'affaires de la grande distribution), ensuite parce que la répartition de leurs points de vente sur l'ensemble du territoire métropolitain n'est vraisemblablement pas parfaitement représentative de la géographie des lieux de consommation des ménages. Au niveau régional néanmoins, les résultats présentés au tableau 3 suggèrent que l'échantillon étudié ne souffre pas de déséquilibre spatial évident par rapport à la distribution de la population.

Par rapport aux travaux existants qui ont été présentés dans la première section, le fait de

mesurer des écarts de niveau de prix conditionnellement à un identifiant univoque du produit – le code-barres en l’occurrence – affermit certainement les constats. De même, l’ensemble des produits pris en compte dans le calcul des différences de niveaux améliore la précision, en raison de leur nombre considérable, et permet de couvrir la quasi-exhaustivité du champ de l’alimentaire et des boissons alcoolisées ou non, référencé par code-barres, tandis que les études antérieures devaient se

contenter de s’appuyer sur des représentants de produits dont la représentativité n’était pas évidente à justifier. Au final, ces travaux apportent d’ores et déjà une information importante et très crédible sur les écarts de niveaux de prix dans l’alimentaire, notamment s’agissant d’agglomérations de taille importante. Ils établissent que la dispersion est relativement faible, comme les travaux historiques l’avaient montré, et qu’elle n’a probablement que peu évolué sur près de 40 ans. □

---

## BIBLIOGRAPHIE

- Anxionnaz, I. & Mothe, A. (2000).** Les comparaisons spatiales de prix au sein du territoire français : historique et développements à prévoir. *Courrier des statistiques*, 98-96, 11–16.  
[https://www.insee.fr/fr/metadonnees/source/fichier/ipc\\_courrierstat\\_95\\_comparaisons\\_spatiales.pdf](https://www.insee.fr/fr/metadonnees/source/fichier/ipc_courrierstat_95_comparaisons_spatiales.pdf)
- Baraille, J. (1978).** Les prix dans les grandes villes de France. *Economie et Statistique*, 106, 17–20.  
[https://www.persee.fr/doc/estat\\_0336-1454\\_1978\\_num\\_106\\_1\\_3004](https://www.persee.fr/doc/estat_0336-1454_1978_num_106_1_3004)
- Baraille, J. P. & Bobin, M. F. (1981).** Les écarts de prix à l’intérieur de la métropole. *Économie et Statistique*, 130, 61–66.  
[https://www.persee.fr/doc/estat\\_0336-1454\\_1981\\_num\\_130\\_1\\_4453](https://www.persee.fr/doc/estat_0336-1454_1981_num_130_1_4453)
- Berthier, J., Lhéritier, J. & Petit, G. (2010).** Comparaison des prix entre la métropole et les DOM en 2010. *Insee Première* N° 1304.  
<https://www.insee.fr/fr/statistiques/1287446>
- Clé, E., Sauvadet, L., Jaluzot, L., Malaval, F. & Rateau, G. (2016).** En 2015, les prix en région parisienne dépassent de 9 % ceux de la province. *Insee Première* N° 1590.  
<https://www.insee.fr/fr/statistiques/1908158>
- de Haan, J. & van der Grient, H. (2011).** Eliminating Chain Drift in Price Indexes Based on Scanner Data. *Journal of Econometrics*, 161(1), 36–46.  
<https://ideas.repec.org/a/eee/econom/v161y2011i1p36-46.html>
- Deaton, A. & Heston, A. (2010).** Understanding PPPs and PPP-based national accounts. *American Economic Journal: Macroeconomics*, 2(4), 1–35.  
<https://doi.org/10.1257/mac.2.4.1>
- Diewert, E. (2003).** Hedonic Regressions: A Consumer Theory Approach. In: Feenstra, R. C. & Shapiro, M. D. (Eds), *Scanner Data and Price Indexes*. Chicago: University of Chicago Press.
- Insee, Division prix de détail (1990).** Les prix dans 23 agglomérations en 1989. *Insee Première* N° 69.  
<https://www.epsilon.insee.fr/jspui/bitstream/1/10075/1/ip69.pdf>
- Fesseau, M., Passeron, V. & Vérone, M. (2008).** Les prix sont plus élevés en Île-de-France qu’en province. *Insee Première* N° 1210.  
<https://www.insee.fr/fr/statistiques/1281287>
- Guglielmetti, F. (1996).** Les prix en Corse : entre Marseille et Paris. *Insee Première* N° 442.  
<https://www.epsilon.insee.fr/jspui/bitstream/1/890/1/ip442.pdf>
- Insee (1998).** Pour comprendre l’indice des prix. *Insee-méthodes* N° 81-82.  
[https://www.insee.fr/fr/metadonnees/source/fichier/Indice\\_des\\_prix.pdf](https://www.insee.fr/fr/metadonnees/source/fichier/Indice_des_prix.pdf)
- Insee, Division prix de détail (1990).** Les prix dans 23 agglomérations en 1989. *Insee Première* N° 69.  
<http://www.epsilon.insee.fr/jspui/bitstream/1/890/1/ip442.pdf>
- Mineau, B. (1987).** Les comparaisons de prix entre agglomérations françaises. *Courrier des statistiques*, 44, 21–24.
- Nicolai, M. P. (2010).** Enquête de comparaison spatiale des prix Corse-Continent 2010. *Quant’île – Insee Corse* N° 12.  
<https://www.insee.fr/fr/statistiques/fichier/1378434/quantile12.pdf>

**Piccard, H. (1972).** Situation relative des prix de détail dans les agglomérations de plus de 20 000 habitants en octobre 1971. *Économie et Statistique*, 37, 35–38.  
<https://doi.org/10.3406/estat.1972.1242>

**Reinsdorf, M. (1999).** Using Scanner Data to Construct CPI Basic Component Indexes. *Journal of Business & Economic Statistics*, 17, 152–160.  
<https://www.jstor.org/stable/1392470>

**Silver, M. & Heravi, S. (2005).** A failure in the measurement of inflation: Results from a hedonic and matched experiment using scanner data. *Journal of Business & Economic Statistics*, 23(3), 269–281.  
<https://doi.org/10.1198/073500104000000343>

**Triplett, J. (2006).** Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes. *Documents de travail de l'OCDE sur la science, la technologie et l'industrie* N° 2004/09.  
<http://dx.doi.org/10.1787/643587187107>

---

