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

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

Abstract – This study is based on scanner data from large retailers sent daily to Insee in 2013. Its aim is to calculate 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 different areas of metropolitan France are essentially structural and change little over time.

JEL Classification: E31, C8, D1
Keywords: price levels, spatial comparison, scanner data

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

Received on 18 September 2017, accepted after revision on 17 January 2019
Translated from the original version: "Écarts spatiaux de niveaux de prix entre régions et villes françaises avec des données de caisse"

To cite this 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

The consumer price monitoring system set up by the French national statistical institute (Insee) essentially aims to determine changes in price levels over time, i.e. inflation. The consumer price index (CPI) is a basis for its measurement. To this end, Insee collectors revisit the same outlets every month to record the prices of the same products, and the overall average change is calculated on the basis of the elementary price changes observed for each product monitored as part of the CPI. Intuition suggests that the price data collected for the purposes of the CPI could also be used to determine average price level differences between different geographical areas of interest. However, this is not generally the case. When measuring average price changes, the aim is to ensure that, when comparing two periods, the same products are actually compared. Similarly, comparing price levels in different geographical areas implies observing the prices of identical products in the areas where price comparisons are conducted. Since this last point, which is specific to the comparison of territorial price levels, is not an issue for the CPI, the product identification process carried out for the purposes of the CPI is generally not detailed enough to ensure that two products observed in two different outlets are identical. In addition, the sample of products tracked in the CPI is obtained by survey and optimised to achieve satisfactory accuracy in measuring inflation at the national level. Shifting to a more granular geographical level automatically raises the problem of the low number of recordings in areas of limited size. Ultimately, even if products were better identified in the CPI, conducting satisfactory comparisons of price levels in different areas would remain a challenge.

Conversely, scanner data are not hampered by some of these limitations for determining spatial price level differences: 1) the barcode (also referred to hereinafter as EAN, standing for European Article Number) is a unique identifier of a product[1]; 2) scanner data cover all transactions relating to industrial food products[2] excluding fresh produce – i.e. fruit, vegetables, shellfish and some fish and meat –, alcoholic and non-alcoholic beverages and some manufactured goods sold in hypermarkets and supermarkets in metropolitan France. The first property referred to above serves to ensure that price comparisons of the same barcode sold in two different stores automatically results in comparing the same product. The second property ensures that the available samples are large enough to allow comparison at a fine level of detail.

Insee initiated a pilot experiment with the aim of integrating scanner data gradually into the calculation of the CPI. To this end, Insee has been receiving daily scanner data from several large retail groups since the end of 2012. The groups involved in the pilot experiment represent approximately 30% of the potential field, i.e. corresponding to the daily transactions of all supermarket chains operating in metropolitan France. The scanner data include, for each store, the list of daily transactions, i.e. the list of barcodes sold, as well as the quantities sold and the corresponding sales prices.[3]

One of the key advantages of scanner data is the wealth of information they provide. The very large volume of data generated means that a far higher level of detail on price levels can be achieved compared to the usual collection system. Scanner data also include both price data and information on the quantity of products sold, thus providing new material for price statistics, which are usually based solely on retail price information. While the first applications naturally concern the determination of inflation at the national level (see, for example, Reinsdorf, 1999; de Haan & van der Grient, 2011), other statistical applications are possible. Comparing price levels across countries remains a complex task since basic products, the product coding system or simply the information systems of supermarket chains are generally not sufficiently alike to allow mass comparisons of EANs. On the other hand, within a single country, where the scanner data information system also provides detailed information according to the place of purchase, scanner data can be used to calculate price level differences between different geographical areas. This is precisely the question examined in this paper, for industrial food products, based on a set of scanner data available to Insee for the year 2013.

Spatial comparison of price levels is a common practice in many countries, usually coordinated by international institutions. Since price levels are bilateral indices, the operation involves defining equivalent classes of products between countries, determining a consumption pattern in

1. In other words, two different products (seen as such by the consumer) necessarily have two different EANs. On the other hand, two different EANs may designate the same product.
2. Unless otherwise stated, the field of industrial food is understood here to mean the field of food products, excluding fresh produce (i.e. fresh fruit and vegetables, shellfish and some fish and meat), and alcoholic and non-alcoholic beverages sold in supermarkets (see the section on Data for more details).
3. In some cases, the corresponding turnover rather than price.

terms of expenditure for the pair of countries considered, identifying products representative of national consumption and comparable in their use in the two countries, and then calculating a bilateral index characteristic of the difference in price levels between countries. One of the main difficulties with this type of operation is to determine classes of products that are genuinely equivalent, i.e. corresponding to an equivalent “use” in the different countries compared. In the absence of the ability to identify identical products – which do not always exist, particularly when countries are relatively different in terms of their cultures and standards of living – the institutions responsible for coordinating such comparisons base the measurement of price differences on comparisons of products with maximally similar characteristics. While this approach provides a good approximation of price level differences based on a compromise between product definition and comparability, it remains open to challenge precisely because of this compromise. The limitations of so-called “purchasing power parity” comparisons are well known and have been detailed in the literature (see, for example, Deaton & Heston, 2010). A key conclusion from this literature is that discussions tend to focus on two different points of limited importance to the comparison exercise conducted here on scanner data and to the task of comparing different French regions. The first point of debate concerns the product comparison exercise, a potentially impossible task when the compared areas differ widely; in this case, the compared areas – i.e. different regions of metropolitan France or conurbations – are very similar in their consumption habits. The second point relates to the method used to calculate indices of level differences. In practice, the methods used generate indices that differ less the closer the prices and consumption structure are in the areas compared.

Of potentially greater importance is the focus of the comparison. By construction, the results presented in this paper relate to the field for which scanner data are available. On the one hand, this means the field of food products (excluding fresh produce) and alcoholic and non-alcoholic beverages sold in supermarkets i.e. industrial food. Therefore, food purchases made in other types of outlets are not included. As such, the results obtained are not representative of food consumption as a whole. In addition, in 2013, Insee only had access to scanner data from a small number of supermarket chains. The corresponding sales represented approximately 30% of supermarket sales in the industrial food sector in metropolitan France. As a result, the regional price level comparisons examined in this paper may be biased because of the choice of supermarkets. The section devoted to presenting the data examines these coverage issues in more detail, showing in particular that the consumption structure obtained from the restricted coverage is consistent with the geographical distribution of the French population. The possible impact of the geographical pricing policy of the major retailers included in the sample is more difficult to determine: if the policy is specific to the retailer and, at the same time, the weight of the retailer in the compared territory differs between the Insee sample and the general picture, all retailers combined, it follows that the index of the territory estimated on the basis of the particular sample will be different from that obtained for all retailers combined. However, on the face of it, the effects of local competition tend to result in price structures becoming standardised across different chains and areas. Therefore, estimates based on a subsample covering 30% of the overall population should, in this context, allow for conclusions of a relatively general nature to be drawn.

The remainder of the paper is organised as follows: the first section presents the results of other comparison exercises aimed at measuring price differences between metropolitan regions and large conurbations carried out by Insee since 1971. The new results obtained from the scanner data used in this study are thus examined in the light of comparable older results. Descriptive statistics are presented in the following section, while a third section presents the model used to analyse the data. The final section presents the results obtained and a robustness analysis, which includes the different discussion points set out above.

Spatial Comparisons at a Metropolitan Territory Level: Some Past Experiences

Studies aimed at comparing price levels between regions of metropolitan France are nothing new since the publications of the General Statistics of France (SGF – late 19th and early 20th centuries) include comparative tables of average retail food prices recorded in different French cities. However, it is only more recently that comparisons have become available that cover a significant range of consumer goods and that are based on a large number of products. Technically, research in this area involves, in the case of comparisons of metropolitan price levels, calculating an average price ratio between the territory concerned and France as a whole for products representative of the consumption of a given variety of products, before aggregating the differences thus measured at the level of product varieties into a national weighted average.[4] The weighting applied in calculating this average corresponds to the national consumption structure, without taking into account local specificities, on the basis that local consumption structures differ very little from the national structure (Mineau, 1987; Anxionnaz & Mothe, 2000). More recent research than the SGF studies includes Piccard (1972) and Baraille (1978), which deal with differences in levels between metropolitan cities. The results of both studies are shown in Table 1. Both studies reach similar conclusions: in the field of food and beverages, the highest food and (alcoholic and non-alcoholic) beverage prices in metropolitan France are found in the Paris conurbation and Corsica. In addition, they show a relatively small dispersion, within a range of slightly less than 10 percentage points.[5] Baraille's study was completed by Baraille & Bobin (1981) using an analysis by type of territory and based on a new survey conducted in 1981. This type of analysis echoed similar results obtained by Piccard (1972).

More recently, Mineau (1987) provided a breakdown by major urban area of differences in food and beverage price levels for 1985; Insee's Retail Price Division (1990) carried out a similar exercise for 1989. The two groups of results show that price level differences between the different areas are stable, as shown in Table 1. Naturally, the two years studied (in this case, 1985 and 1989) are close, although a similar

4. *With the notable exception of the most recent studies on spatial price comparisons based on ad hoc surveys (Nicolaï, 2010; Berthier* et al., *2010; Clé* et al., *2016). These studies are based on an approach inspired by harmonised European surveys conducted to measure purchasing power parities and use Fisher price indexes, based on consumption patterns specific to each of the territories compared. This approach is justified when consumption patterns differ significantly between the territories compared, as is the case, for example, between French overseas departments and metropolitan France. On the other hand, when comparing different regions of metropolitan France, differences in regional structures tend to be very limited and taking them into account is a secondary issue.*

5. *Baraille (1978) study measured an 8% gap between the prices of food and beverages in the urban area where they were the highest (Ajaccio-Bastia) and the lowest (Angers).*

Table 1
Average price differences observed in metropolitan France in the food and beverage sector

Area	Index, from the results of :				
	Piccard (1972) year 1971	Baraille (1978) year 1977	Mineau (1987) year 1985	Insee (1990) year 1989	Guglielmetti (1996) year 1995
Paris conurbation	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	

Notes: The overall level of the indices is set with reference to the Paris conurbation (recalculated by the authors from the data published for reference to the Paris conurbation).

result applies to 1977, which is more distant. In these studies, we see once again that food and beverage prices are higher in Corsica than anywhere else. The Paris conurbation, where consumer prices are 2 to 3% higher than in provincial cities, comes second.

The study for 1995 by Guglielmetti (1996) found that the average difference in the level of prices for food and beverages (alcoholic and non-alcoholic beverages, including tobacco) in Corsica was significantly higher than in 1989, reaching 8.5% compared to Paris, with the gap between Paris and Marseille remaining unchanged over the period.

The results of the most recent and more widely applicable studies carried out do not deviate significantly from these findings. Fesseau *et al.* (2008) found that food and non-alcoholic beverage prices were approximately 5.7% higher in Île-de-France than in the provinces in 2006. Based on a spatial comparison survey of price levels carried out by Insee in 2010, Nicolaï (2010) established that the average price levels of food and non-alcoholic beverages were approximately 8.6% higher in Corsica than on the continent as a whole. Finally, the same survey conducted in 2015 showed that food[6] and non-alcoholic beverage prices in that year were 6.5% higher in the Paris region than in the provinces and 2.1% higher in Corsica than in the Paris region (Clé *et al.,* 2016). Therefore, these latter results, based on data collected to measure price level differences, confirm the hierarchy and orders of magnitude previously established for the food sector.

Ultimately, these various studies, the scope, methodology and nature of which differ somewhat, provide broadly consistent results: differences in price levels are highly structural characteristics, meaning that they change relatively little over time; prices are higher in Corsica, probably because it is an island, which limits competition and increases production costs, notably on account of the transport costs of products produced on the continent; they are also higher in the Paris region, probably because of higher marketing costs (commercial property prices) and the purchasing power of resident consumers, which is on average higher than elsewhere.

The Data

The data used are the scanner data of distribution chains that have entered into an agreement authorising Insee to access daily records for 2013. Within these data, only those related to industrial food, i.e. food products and beverages (both alcoholic and non-alcoholic[7]) sold in supermarkets, are included in the study. The data were obtained from 1,833 stores in April 2013. The stores are located in 1,330 municipalities in 707 urban areas of metropolitan France.[8] The distribution of the number of outlets in the major urban areas included in the studies referred to earlier is given in Table 2.

The distribution by region is shown in Table 3. Note that these are, here as in the entire article, the administrative regions prior to the 2015 reform (NOTRe Act). Overall, the distribution of the number of outlets at the regional level is relatively similar to the demographic distribution. In other words, because of their geographical distribution, the outlets included

6. Also including fresh produce.
7. Division of COICOP 01, excluding fresh produce (fresh fruit and vegetables, shellfish and some fish and meat) and Group in COICOP 02.1.
8. Classification of Urban Units, 2010 version. The classification includes around 2,000 units.

Table 2
Number of retail outlets per large urban area in the sample

Urban Area	Number of retail outlets
Paris conurbation	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

Notes: When the number of points of sale is less than or equal to 4 (Limoges, Dijon, Poitiers, Ajaccio-Bastia), the city index does not appear in the results table (see Table 7).
Sources: Insee, scanner data 2013.

in the database provide a relatively accurate picture of the French retail landscape. Naturally, insofar as only a limited number of large retail groups submitted their data to Insee in 2013, cluster effects remain to be feared.

The consumption structure in terms of products consumed should theoretically be similar from region to region. To examine this hypothesis, we calculated the structure using the scanner database. Table 3 shows the breakdown of turnover associated with product groupings according to the Classification of Individual Consumption by Purpose (COICOP). As expected, the statistics show that regional structures in the industrial food sector differ little from the average metropolitan structure relating to the same coverage. It should also be noted that this structure, which is specific to purchases made in supermarkets, differs significantly from the consumption structure for all forms of sales combined, mainly for non-industrial fresh produce (fresh fruit and vegetables, shellfish, some fish and meat).

Thus constructed, the database includes, on average, 16.4 million observations per week, corresponding to the intersection [outlet × EAN] of average prices per barcode and turnover. The total turnover for a week of observation available in the database stands, on average, at around €445 million. Extrapolated over a year (52 weeks) and related to household consumption expenditure[9] recorded in 2012 and spent on food and alcoholic and non-alcoholic beverages, this turnover figure represents approximately 15% of the consumption expenditure of households within the field of study.[10]

9. National Accounts report 156 billion euros (current euros).

10. To be precise, the differences within the field relate to food products sold in other outlets (of major retailers among the supermarkets not included in the study because they did not send their data to Insee in 2013, as well as other types of stores or markets) and to fresh produce.

Table 3
Number of retail outlets per region in the sample

Region	Number of retail outlets	Weights (in %)	Demographic weight (in %)
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

Reading note: In the data used, the Île-de-France region includes 404 points of sale. The 404 outlets represent 22.1% of the 1,829 outlets in the database. As a reminder and comparison, the Île-de-France region represents 18.8% of the inhabitants of metropolitan France (Population Census, 2012). The figures in italics are not from the scanner database.
Sources: Insee, scanner data 2013.

Estimation Model

A single observation corresponds to a barcode (EAN) sold in a store in the sample during the week under consideration. In other words, one observation per [outlet × EAN] is recorded. It is assumed that the single observations thus defined are identified by index i of set I. Thus, p_i is the price (unit value over the week) of the item identified by its barcode in one of the stores included in the database. Let ω_i be the turnover associated with the corresponding observation.

The index reflecting price level differences between geographical areas is calculated using a hedonic method (Triplett, 2006). This approach, based on econometric price modelling, differs somewhat from the harmonised approaches used to measure purchasing power parities across European countries. Nevertheless, it is known as one of the traditional methods (Deaton & Heston, 2010) and, where the territories compared present similar consumption patterns (in terms of price and structure, as is the case here – see Table 4), it results in price level differences similar to those found using alternative methods.

The econometric model is based on the barcode and the geographical area of origin of product i considered. By using the barcodes, the model used allows for the average price differences between geographical areas to be estimated.

Table 4
Regional consumption structures in the field of industrial food

Region	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
Metropolitan France (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

Notes: Territorial distribution (%) of turnover, according to product type, by grouping classes of the COICOP nomenclature. 01.1.1: Bread and cereals; 01.1.2: Meat; 01.1.3: Fish and shellfish; 01.1.4: Milk, cheese and eggs; 01.1.5: Oil and fat; 01.1.6: Fruit; 01.1.7 Vegetables; 01.1.8: Sugar, jams, chocolate, confectionery and iced products; 01.1.9: Salt, spices, sauces and food products not elsewhere; 01.2.1: Coffee, tea and cocoa; 01.2.2: Other soft drinks; 02.1.1: Alcohols; 02.1.2: Wines, cider and champagne; 02.1.3: Beers.
Calculation by the authors based on the fund data for the reference week (April 2013), including for (1). (2) Breakdown by country, National Accounts (detailed household consumption tables for 2013).
Sources: Insee, scanner data 2013.

Formally, it is assumed that price p_i responds to a generating process of the form:

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

where **1** denotes a dummy variable equal to 1 if the condition in parentheses in index is true and 0 if not, ean_i is the barcode number of observation i and $zone_i$ is the geographical area to which observation i relates. ε_i is a centred random variable. In this model, coefficients c, α_ℓ ($\ell \in \{1, \dots ,L\}$, L is the number of barcodes taken into account) and β_z ($z \in \{1, \dots ,Z\}$, Z is the number of geographical areas taken into account) are not known. They are estimated by least squares. The ratios[11] of coefficients α_ℓ are interpreted as the average price ratios associated with the barcodes considered. The ratios of coefficients β_z reflect the average price ratios between geographical areas for given products (identified by their barcodes). These coefficients, estimated by least squares, correspond to hedonic price indexes (Triplett, 2006; Diewert, 2003; Silver & Heravi, 2005).

The form of the estimators obtained is detailed in the Box below. We see that the resulting estimator naturally takes into account the differences in consumption structure between regions through the weightings used. From this point of view, the most natural weighting is by the turnover of the product in the outlet considered. Therefore, the reference model involves a weighting by turnover. Unit weighting *de facto* involves a structure relatively similar to that of turnover since it is based on the number of transactions for the product and outlet in question. The alternative approach by unit weighting is therefore used to examine the robustness of the results with respect to the reference weighting.

11. *To be precise, the exponential ratio of these coefficients (see* infra*).*

Box – **Structure of Hedonic Estimators**

The least squares estimator (1) may or may not be weighted. In practice, there are two possible options: either we use weightings similar to the turnover figures ω_i, or single observations are not weighted. To properly assess the consequences of the choice of weightings, it is useful to examine the structure of the estimators we obtain for the β_z coefficients. To this end, we assume, for greater simplicity, that the estimation is carried out in two steps[(a)]: a first step in which the α_ℓ coefficients are estimated; then, in a second step, the β_z coefficients are estimated (conditional on the estimators $\hat{\alpha}_\ell$ of the α_ℓ obtained in the first step). Of course, by proceeding in this way, we do not obtain the same least squares estimator that we would if the vectors ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) were estimated simultaneously, but the probability limits of the two two-step estimators are the same as those of the one-step estimator[(b)]. The advantage of proceeding in two steps is that it is easy to derive the form of $\hat{\beta}$. Let $\tilde{p}_i$ be the adjusted variable p_i of the first step:

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

The second step consists in regressing $\log(\tilde{p}_i)$ on the vector line $\boldsymbol{x}_i$ comprising Z columns, of which $Z-1$ are null, and the only non-zero column is equal to 1:

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

The least squares estimator $\hat{\boldsymbol{\beta}}$ is traditionally the solution of the equation (here in a weighted version; for an unweighted version, simply let $\omega_i = 1$):

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

where, by grouping by zone mode[(c)]:

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

and lastly, for the zone k considered (also expressed as z_k):

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

It follows that, for zones k and j, we have:

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

It should be noted that this ratio corresponds to an average price ratio[(d)] (i.e. unit value ratio). We find that the index of price level differences between zones takes into account local consumption patterns since, in both the numerator and the denominator, the weight of each product in the index is in proportion to its weight in local consumption expenditure.

(a) This two-step decomposition is given merely to clarify the form of the resulting index. In practice, a one-step calculation is performed based on model (1).
(b) Under the same convergence assumptions, including orthogonality of the random variable and explanatory variables.
(c) Diag denotes the diagonal matrix whose diagonal coincides with the vector as an argument.
(d) As a geometric mean, to be interpreted as being calculated with a fixed barcode, identical for the numerator and denominator, due to the conditioning by the EAN in steps (1) and (2).

At this stage, the conditions under which the estimator $\hat{\beta}_k$ is unbiased need to be specified. As an estimator of coefficient β_k in equation (1), the coefficient is unbiased when the orthogonality conditions of the explanatory variables and random variable ε_i (or v_i in the case of the second-step regression) are satisfied. It is assumed here that this is the case. On the other hand, statistics $\exp(\hat{\beta}_k)$ are not an unbiased estimator of $\exp(\beta_k)$. Indeed, based on the expression of the least squares estimator (equation 1 or 3), we show[12] that:

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

where σ^2 is the variance of the ε_i, which will now be assumed to have the same variance. This correction will therefore be used to calculate the price ratios.

Results

Differences Observed in April 2013

This section presents the results based on regressions similar to the regression of model (1) for a week of data in April 2013 (third week of the month). In practical terms, for all the regressions performed, only 5,000 barcode references per supermarket chain are retained. Among the supermarket chain's sold references, the top 5,000 in terms of turnover are retained. Hedonic model (1) is based on barcode dummies. These are not explicitly estimated (they are reduced algebraically in the normal equation), but too many references produce a normal equation that is too complicated to process. Various tests were conducted to examine the consequences of this restriction. The tests show that retaining 3,000 or 5,000 references per supermarket chain does not lead to significantly different results based on geographical dummies. Ultimately, the combination, for all the supermarket chains included in the base, of the 5,000 main barcodes relating to them, leads to considering 13,098 barcodes in the regressions. This number is significantly higher than the 5,000 references retained per supermarket chain, meaning that a significant proportion of barcodes are specific[13] to supermarket chains (own brands). Given this restriction, the basis of calculation includes 7.3 million records corresponding to the intersections [outlet × barcodes] retained. In terms of turnover, the restriction applied results in 74% of the information contained in the original database presented in the section on Data being retained.

Table 5 shows, by product type and for the database restricted to the 5,000 main barcodes per supermarket chain, the number of IRI[14] product families associated with them, as well as the corresponding number of barcode references. Roughly speaking, an IRI family corresponds to a type of product approximately as fine grained as the varieties of products tracked by the CPI (Insee, 1998). As a reminder, 327 varieties were tracked in the metropolitan CPI for industrial food in 2013. This figure is comparable to the number of IRI families which, based on the same coverage, totals 288. In the database studied, the corresponding number of barcode references stands, as noted above, at 13,098.

Table 6 shows the estimation results of the gap in price indices level for industrial food in administrative regions of metropolitan France, calculated using the scanner database. First, we see that the dispersion of the differences is relatively small: 5.5 to 8 percentage points depending on whether or not the observations are weighted by their turnover. The dispersion is greater when considering unweighted indices rather than weighted indices, suggesting that products with a greater weight in consumer budgets have a lower spatial price dispersion than other products. It is also worth noting that the ranking of the regions by average price difference level remains unchanged whether or not the observations are weighted by turnover.

Geographically, the results highlight distinct regional trends: a large central-west region of France where price levels are approximately 3% lower than in Île-de-France; then a category that includes the more rural regions of central France, those of northern France and Aquitaine where industrial food prices are on average 2% lower than in Île-de-France; and the more industrial and urban regions of eastern and southern France have food price levels 1% lower than in Île-de-France. Lastly, prices in Corsica are 2% higher than in the Île-de-France region.

In order to compare the "historical" results shown in Table 1, Table 7 groups the indices of industrial food price differentials between the major metropolitan areas and the Paris conurbation. When comparing these results

12. For example, by using a Δ-method or by making assumptions about the normal distribution of random variables in equation (1). E *stands for the mathematical expectation (conditional notation).*

13. If each barcode was sold in all stores, the combination of the 5,000 main store barcodes would include precisely 5,000 barcodes.

14. Private firm that develops a catalogue (used by Insee as part of the pilot experiment) of characteristics of products indexed by barcodes.

Table 5
Distribution of IRI families and barcodes by COICOP nomenclature grouping

COICOP code	COICOP - product	Number of families	Number of barcodes
0111	Bread	47	2,200
0112	Meat	19	1,479
0113	Fish	22	848
0114	Milk, cheese, eggs	23	1,830
0115	Oil and fat	6	300
0116	Fruits	15	252
0117	Vegetables	31	1,117
0118	Sugar, preserves, chocolat, sweets, icecream	29	1,098
0119	Salt, spices, sauces and other	35	564
0121	Coffee, tea, cocoa	10	409
0122	Other non-alcoholic beverages	17	876
0211	Alcohol	12	361
0212	Wine, cider, champaign	21	1,535
0213	Beers	1	229
Total		288	13,098

Reading note: The database includes 47 IRI product families belonging to the COICOP 0111 grouping (Breads). 2,200 barcode references refer to it in the database examined.
Sources: Insee, scanner data 2013.

Table 6
Price level gap indices between the Paris region and other regions

Region	Code	Estimation		
		Weighted	Unweighted	Weighted with retail E.F.
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 (Ref.)**	100 (Ref.)	100 (Ref.)
Corse	94	**102.1**	103.5	102.8

Reading note: According to the estimate in which the observations are weighted by their turnover, prices are on average 3.3% lower in Brittany than in the Île-de-France region. According to the estimate in which the observations are weighted individually, prices are on average 4.4% lower in Brittany than in Ile-de-France. The zone indicators result from a regression of type (1) in which the zones are the former administrative regions. The last column refers to a calculation equivalent to that made for the first column (i. e. weighted), in which a fixed effect has been added. The results obtained are corrected according to formula (6) and transformed into indices by a multiplication by 100. The estimated variance of the hazard is 0.004. Calculation based on 7.3 million records. The average standard deviation on the indices presented is 0.02 index points.
Sources: Insee, scanner data 2013.

with the results shown in Table 1, it should be recalled that the economic and geographical coverage and the calculation methods used are not strictly identical. Some of the differences found between conurbations and their variation over time probably include biases due to inconsistent coverage and methods. Nevertheless, the results obtained are still worth examining.

For both conurbations and regions, the findings show (see Tables 6 and 7) that the differences in price levels estimated by unweighted regression are slightly greater than those calculated using weighted regression. Excluding Corsica[15], price differences are in the range of 3.7 to 4.4 percentage points depending on whether or not the observations are weighted. Compared to the Paris conurbation, where prices are highest, the least expensive conurbations (among the major conurbations) for industrial food are Nantes, Rennes, Orléans, Rouen and Lille. Remarkably, this was also the case in 1989 (Insee, Retail Price Division 1990) and 1985 (Mineau, 1987) – see Table 1. The difference with the 1977 picture (Baraille, 1978) is slightly greater.

Compared to the differences between regions, the differences found between large conurbations are slightly more pronounced. For example, with reference to an almost comparable area (the Paris conurbation or the Île-de-France region as the case may be), the (weighted) index for Montpellier is 97.9 while that for Languedoc-Roussillon is 98.6. Similarly, the index for Lille is 97.3 while the index for Nord-Pas-de-Calais is 97.9.

15. *Not presented in Table 7 because of the excessively small number of outlets in the scanner database.*

Table 7
Price level differences between the Paris metropolitan area and the main other metropolitan areas

Area	Estimation	
	Weighted	Unweighted
Paris conurbation	100 (Ref.)	100 (Ref.)
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	n.a.	n.a.
Ajaccio-Bastia	n.a.	n.a.
Clermont-Ferrand	98.4	97.3
Toulouse	98.0	96.7
Dijon	n.a.	n.a.
Nantes	96.3	95.9
Nancy	98.4	97.7
Poitiers	n.a.	n.a.
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

Reading note: According to the estimation in which the observations are weighted by their turnover, prices are on average 1.4% lower in Lyon than in Paris. According to the estimate in which the observations are weighted individually, prices are on average 2.3% lower in Lyon than in Paris. The zone indicators result from a type (1) regression in which the zones are agglomerations (urban units). The results obtained are corrected according to formula (6) and transformed into indices by a multiplication by 100. The estimated variance of the hazard is 0.004. Calculation on 7.3 million records. The average standard deviation on the indices presented is 0.10 index points.
Sources: Insee, scanner data 2013.

This may be related to the fact that competition is probably greater in local markets in large conurbations, which tends to drive prices down.[16]

However, there are two exceptions to this rule among large conurbations: Strasbourg, which has an index of 99.1, compared to 98.9 for Alsace, and Clermont-Ferrand with an index of 98.4, compared to 98.2 for Auvergne. In both cases, however, the differences are not significant.

As noted in the introduction and above, the representativeness of the data sample in relation to the spatial distribution of prices may be affected because of the limited number of supermarket chains that provided their data to Insee in 2013. Thus, the selection of the supermarket chains included in the sample may be correlated to the regional dimension on which the proposed statistics are estimated. This is the case, for example, if a supermarket chain included in the sample with a pricing policy different from the other chains (for example if its prices are invariably lower) is, as a result of the selection, over-represented in one region and not elsewhere. In this case, the estimation of the price level in the over-represented region is biased (downward in the case of the example given) compared to other regions.

A complete dataset for all supermarket chains would be required to demonstrate whether or not such a bias exists and to assess it. While it is not possible to carry out a definitively conclusive study on this point based on the limited sample available, it is possible to examine whether some of the results are consistent with the assumption of representativeness of the subsample used. The first finding of interest in this regard was presented in Table 3, which shows that the regional distribution of outlets is consistent with the distribution of the population and therefore, probably, with household food consumption expenditure. Another finding of interest is to add supermarket chain dummies to equation (1). If, for example, a regional index is significantly different in the second calculation compared to the reference calculation, the implication is that the regional price level is partly explained by the chains represented in the local supermarket network in the subsample used. Given this, it may be that the results obtained are essentially limited in scope to the sample considered. A calculation along these lines was carried out, the results of which, in terms of regional indices weighted by sales, are shown in the last column of Table 6. The results can be compared to those of the reference calculation (in bold in Table 6). It appears that the regional indices can be quite significantly different, up to 0.8 points in the case of Bourgogne and Franche-Comté. However, the main findings, particularly as regards the price hierarchy between Corsica, Île-de-France and other metropolitan regions, as well as the order of magnitude of the differences, remain true.

Finally, if "supermarket chain effects" clearly exist, with their impact on local indices being visible, the various robustness tests carried out provide some evidence that the main lessons learned from the subsample are reasonably substantiated for the whole of food consumption in the supermarket sector.

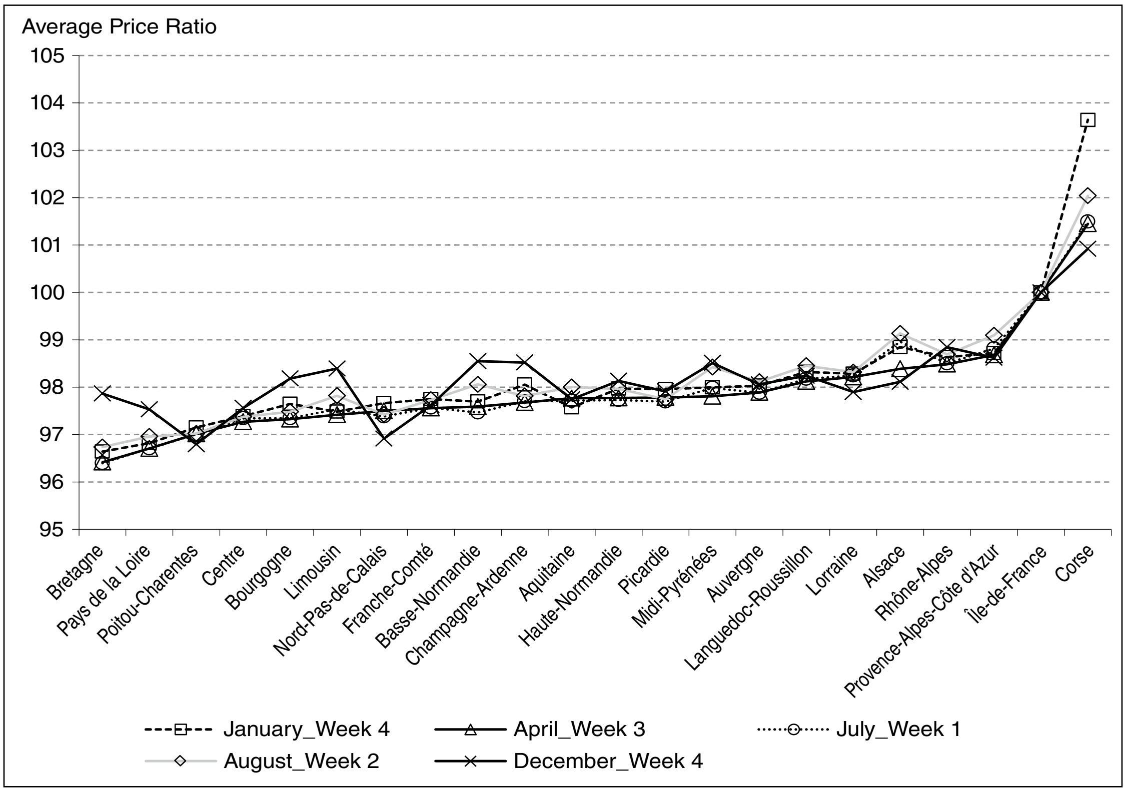
Sensitivity of the Results to the Choice of Study Week

To test the robustness of the results obtained, we examine here how regional differences in price levels behave when selecting a different study week. To do this, the analysis is extended to four other weeks in 2013 that are relatively typical in terms of sales and holidays with a strong impact on consumer purchases: the fourth week of January (shortly after the Christmas and New Year festivities and in the middle of the winter sales period), the first week of July (beginning of the summer holidays), the second week of August (end of the summer holidays) and the fourth week of December (Christmas and New Year festivities). The selected weeks are compared to the third week of April studied above and used as a reference for comparison.

The following Figure shows price level differences between the Île-de-France region and the other regions for the 5 weeks studied, the regions being ordered on the x-axis according to their rank in terms of the price level observed during the April reference week. The results show that the gaps are very close from one week to the next, with two exceptions. First, price levels in Corsica are higher in January compared to the other weeks studied. Second, we found a relatively significant change in the regional price structure during the last week of December, interpreted as the likely effect of the specific nature of the products sold at that time and the large population movements during the holidays, which alter the geographical structure of the markets.

16. For interpretation purposes, we make the assumption (a reasonable assumption given its weight) that the price level of the Paris conurbation is also the price level of the Île-de-France region. Consequently, the differences in the indices of provincial cities and their regions are assumed to be linked to local differences between the cities and their regions and not to possible price differences between the urban unit of Paris and its region.

Figure
Difference in regional price levels compared to Île-de-France



Notes: Reference 100 for Île-de-France for each week of study. The regions shown on the x-axis are in increasing order according to the index level recorded in April 2013.
Sources: Insee, scanner data 2013.

Ultimately, the robustness analysis tends to confirm the broadly structural nature of geographical differences in price levels. It also demonstrates the richness of scanner datasets as a means of accurately estimating price indices over geographical or temporal ranges inaccessible to traditional survey methods.

* *
*

This study provides an example of how scanner data can be used to measure price level differences between areas of metropolitan France in the field of food and alcoholic and non-alcoholic beverages. Naturally, given the nature of the scanner data used, the results remain limited in scope and extending them to all food consumption by metropolitan households is clearly open to discussion. The first reason for this is that only a relatively small number of supermarket chains participated in the pilot experiment conducted by Insee in 2013 (despite accounting for 30% of the turnover of the major supermarkets); the second reason is that the distribution of their outlets across the territory of metropolitan France is probably not perfectly representative of the geography of household consumption sites. At the regional level, however, the results shown in Table 3 suggest that the study sample does not suffer from an obvious spatial bias with respect to the distribution of the population.

Compared to the previous research discussed in the first section, measuring price level differences conditional on a unique product identifier – in this case, a barcode – clearly reinforces the findings. Similarly, all the products taken into account in calculating differences in levels serve to improve accuracy because of their considerable volume and allow for an almost exhaustive coverage of all food products and alcoholic and non-alcoholic beverages, referenced by barcodes, while previous studies were forced to rely only on representatives of products whose representativeness was difficult to prove. Ultimately, this study provides important and highly credible information on spatial differences in food price levels, particularly in the case of large urban areas. The findings demonstrate that the dispersion is relatively low, as historical research has shown, and that it has probably changed very little over nearly 40 years. ☐

BIBLIOGRAPHY

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

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

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

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 Consummer 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

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.

Nicolaï, 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