Combining micro and macro data in hedonic price indexes

Esmeralda A. Ramalho and Joaquim J.S. Ramalho

Department of Economics and *CEFAGE, Universidade de Évora*

February 2015

Abstract

This paper proposes arithmetic and geometric Paasche quality-adjusted price indexes that combine micro data from the base period with macro data on the averages of asset prices and characteristics at the index period. The suggested index has two types of advantages relative to traditional Paasche indexes: (i) simplification and cost reduction of data acquisition and manipulation; and (ii) potentially greater efficiency and robustness to sampling problems.

Keywords: Paasche price index, imputation hedonic method, quality adjustment, asset heterogeneity.

1 Introduction

Hedonic methods are a prominent approach in the construction of quality-adjusted price indexes (QAPI) for infrequently traded heterogeneous assets such as houses (see, *e.g.*, Hill and Melser, 2008), artworks (*e.g.*, Collins, Scorcu and Zanola, 2009) and collectables (*e.g.*, Georges and Seçkin, 2012). All hedonic methods require the estimation of a regression equation relating asset prices to asset characteristics. The parameters of this so-called hedonic function provide a measure of the implicit marginal price of each asset characteristic and therefore this function may be used to predict the asset prices at different time periods while controlling for their heterogeneity.

The most common and flexible hedonic method is the imputation price method, which allows the implicit prices of the asset characteristics to vary freely over time. In general, QAPI based on this method require the estimation of an hedonic function *at each time period*. However, several authors (*e.g.*, Pakes, 2003) have shown that it is possible to compute arithmetic and geometric Paasche QAPI by estimating the hedonic function only *at the base period*, although a sample of micro data on asset prices and characteristics still needs to be collected *for all periods*. The main aim of this paper is to show that, actually, a sample of micro data needs to be collected also *only for the base period*. For the other periods, it is enough to use aggregate information about asset prices and characteristics, namely their arithmetic or geometric averages, which may arise from the same source used for the base period or from any other source.

The suggested Paasche QAPI that combines micro and macro data has several advantages relative to the corresponding index that uses only micro information. On the one hand, the strong micro data requirements that characterize the hedonic approach are restricted to the base period. Thus, the data acquisition and preparation process is simplified and more cost-effective. Indeed, aggregate data does not raise confidentiality issues and often are substantially cheaper than individual data or even publicly available. Moreover, macro data can be directly combined

in the index formula, avoiding the complex matching processes usually required to merge micro information released by different sources. On the other hand, because the aggregate information may be obtained from larger samples or even the whole population of interest, displaying little or no sampling error (see Imbens and Lancaster, 1994), its inclusion in the index formula produces precision gains and reinforces the index robustness to various sampling problems that commonly affect micro data, such as missing data and measurement error.

2 Paasche quality-adjusted price indexes

Let p_{it} be the price p of asset i at period t, where i indexes different assets at each time period. We assume that either $t = 0$ (base period) or $t = s$ (current period). Let N_t be the number of assets observed at each period. Let $X_{it,j}$ be (a function of) the characteristic j of asset i at period t, $j = 1, ..., k$, and let x_{it} be the $1 \times (k+1)$ vector with elements $X_{it,j}$, $j = 0, ..., k$, where $X_{it,0} = 1$ denotes the constant term of the hedonic regression. Let $\bar{X}_{t,j} = N_t^{-1} \sum_{i=1}^{N_t} X_{it,j}$ and denote by $\bar{x}_t$ the $(k+1)$-vector containing the sample averages of the asset characteristics. Finally, let the superscript $R = \{A, G\}$ denote a quantity associated to an arithmetic (A) or geometric (G) index.

2.1 Traditional calculation

The unadjusted, fixed base arithmetic and geometric price indexes for period s for infrequently traded heterogeneous assets are defined, respectively, by the following ratios:

$$I_s^A = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s} p_{is}}{\frac{1}{N_0}\sum_{i=1}^{N_0} p_{i0}} \qquad \text{and} \qquad I_s^G = \frac{\prod_{i=1}^{N_s} p_{is}^{\frac{1}{N_s}}}{\prod_{i=1}^{N_0} p_{i0}^{\frac{1}{N_0}}} = \frac{\exp\left[\frac{1}{N_s}\sum_{i=1}^{N_s} \ln(p_{is})\right]}{\exp\left[\frac{1}{N_0}\sum_{i=1}^{N_0} \ln(p_{i0})\right]}. \tag{1}$$

As shown by Reis and Santos Silva (2006), for each index it is particularly appropriate to use hedonic functions where the scale of the price corresponds to that of the index. Otherwise, complex retransformation bias corrections have to be estimated to obtain consistent estimators for I_s^R; see Ramalho and Ramalho (2014). Thus, for constructing an estimator for I_s^A (I_s^G), we consider only hedonic functions that use the price (logged price) as dependent variable. In this paper we assume additionally that the hedonic function is linear in the parameters, being written as $p_{it} = x_{it}\beta_t^A + u_{it}^A$ (arithmetic indexes) or $\ln p_{it} = x_{it}\beta_t^G + u_{it}^G$ (geometric indexes), where u_{it}^R is an error term and β_t^R is a vector of parameters with elements $\beta_{t,j}^R$. The parameter $\beta_{t,j}^R$ is often interpreted as the implicit marginal price for the asset characteristic $X_{it,j}$.

After estimating the hedonic functions for both the base and current periods, consistent estimators for I_s^A and I_s^G are given by, respectively,

$$\hat{I}_s^A = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s} \widehat{p_{is}}}{\frac{1}{N_0}\sum_{i=1}^{N_0} \widehat{p_{i0}}} = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_s^A}{\frac{1}{N_0}\sum_{i=1}^{N_0} x_{i0}\hat{\beta}_0^A}$$

and

$$\hat{I}_s^G = \frac{\exp\left[\frac{1}{N_s}\sum_{i=1}^{N_s} \widehat{\ln(p_{is})}\right]}{\exp\left[\frac{1}{N_0}\sum_{i=1}^{N_0} \widehat{\ln(p_{i0})}\right]} = \frac{\exp\left(\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_s^G\right)}{\exp\left(\frac{1}{N_0}\sum_{i=1}^{N_0} x_{i0}\hat{\beta}_0^G\right)}.$$

Both estimators may be straightforwardly decomposed into a QAPI ($\hat{I}_s^{R_p}$) and a quality index ($\hat{I}_s^{R_q}$): $\hat{I}_s^R = \hat{I}_s^{R_p}\hat{I}_s^{R_q}$, where:

$$\hat{I}_s^{A_p} = \frac{\frac{1}{N_a}\sum_{i=1}^{N_a} x_{ia}\hat{\beta}_s^A}{\frac{1}{N_a}\sum_{i=1}^{N_a} x_{ia}\hat{\beta}_0^A}, \qquad \hat{I}_s^{G_p} = \frac{\exp\left(\frac{1}{N_a}\sum_{i=1}^{N_a} x_{ia}\hat{\beta}_s^G\right)}{\exp\left(\frac{1}{N_a}\sum_{i=1}^{N_a} x_{ia}\hat{\beta}_0^G\right)}, \tag{2}$$

$$\hat{I}_s^{A_q} = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_b^A}{\frac{1}{N_0}\sum_{i=1}^{N_0} x_{i0}\hat{\beta}_b^A}, \qquad \hat{I}_s^{G_q} = \frac{\exp\left(\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_b^G\right)}{\exp\left(\frac{1}{N_0}\sum_{i=1}^{N_0} x_{i0}\hat{\beta}_b^G\right)} \tag{3}$$

and $(a,b) = (0,s)$ ($\hat{I}_s^{R_p}$ is a Laspeyres index) or $(a,b) = (s,0)$ ($\hat{I}_s^{R_q}$ is a Paasche index). While $\hat{I}_s^R$ is an estimate of the overall asset price change between periods 0 and s, $\hat{I}_s^{R_p}$ measures only pure price movements (the same asset characteristics are used in the numerator and denominator of equation 2) and $\hat{I}_s^{R_q}$ measures only quality changes (the implicit prices of the asset characteristics are fixed at $\hat{\beta}_b^R$ in the calculation of the index).

The most common way of calculating a hedonic QAPI is through direct application of the formulas in (2). However, in the case of Paasche indexes a very convenient simplification applies, provided that the hedonic function is estimated by ordinary least squares (OLS). Indeed, because the sum of OLS residuals is zero by definition and hedonic functions typically include an intercept term, it follows that $\sum_{i=1}^{N_t} p_{it} = \sum_{i=1}^{N_t} \widehat{p_{it}} = \sum_{i=1}^{N_s} x_{it}\hat{\beta}_t^A$ (arithmetic indexes) and $\sum_{i=1}^{N_t} \ln(p_{it}) = \sum_{i=1}^{N_t} \widehat{\ln(p_{it})} = \sum_{i=1}^{N_s} x_{it}\hat{\beta}_t^G$ (geometric indexes). Hence, (2) may be simplified to:

$$\hat{I}_s^{A_p} = \frac{\bar{p}_s^A}{\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_0^A} \qquad \text{and} \qquad \hat{I}_s^{G_p} = \frac{\bar{p}_s^G}{\exp\left(\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_0^G\right)}, \tag{4}$$

where $\bar{p}_s^R$ denotes the sample arithmetic or geometric mean of the asset prices in the current period. Hence, unlike suggested by equation (2), the hedonic function needs to be estimated only at the base period.

2.2 Combining micro and macro data

Although simpler than (2), the Paasche QAPI expressed in (4) still require individual asset data for all periods. However, equation (4) may be further simplified in order to express $\hat{I}_s^{R_p}$ as a function of only $\hat{\beta}_0^R$ and aggregate data. In fact, given that $N_s^{-1}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_0^R = N_s^{-1}\sum_{i=1}^{N_s}\sum_{j=0}^{k} X_{is,j}\hat{\beta}_{0,j}^R = \sum_{j=0}^{k} \bar{X}_{s,j}\hat{\beta}_{0,j}^R = \bar{x}_s\hat{\beta}_0^R$, it follows that (4) may be written as:

$$\hat{I}_s^{A_p} = \frac{\bar{p}_s^A}{\bar{x}_s\hat{\beta}_0^A} \qquad \text{and} \qquad \hat{I}_s^{G_p} = \frac{\bar{p}_s^G}{\exp\left(\bar{x}_s\hat{\beta}_0^G\right)}. \tag{5}$$

The aggregation of the characteristics across assets at the current period completely removed the direct dependence of the index formula on micro data for period s with *no loss of information*, because (5) is numerically equal to both (2) and (4).[1]

Consistency of the Paasche QAPI in (5) requires consistent estimation of: (i) the implicit prices of all relevant characteristics at the base period; and (ii) the means of the asset prices and characteristics at the current period.[2] While the implicit prices β_0^R have to be estimated from a micro dataset containing asset prices and all relevant asset characteristics, the averages $\bar{p}_s^R$ and $\bar{x}_s$ do not have to be necessarily estimated from the corresponding micro dataset for period s. In

[1] Although not exploited in this paper, the same technique may be directly applied to the measurement of quality/productivity changes, since the quality indexes in (3) may be written as $\hat{I}_s^{R_q} = \exp\left(\bar{x}_s\hat{\beta}_0\right)/\bar{p}_0^R$. In this case, it is not even required any type of information on asset prices for the index period.

[2] Note that only the means of the characteristics that are relevant at the base period are necessary.

fact, $\bar{p}_s^R$ and $\bar{x}_s$ may be directly obtained in the form of aggregate information, which may come from other sources that use larger samples of micro data but either do not release individual data or provide them at a high cost. In certain cases, those larger samples may even coincide with the population of interest. Define these aggregate quantities arising from other sources as $\bar{p}_s^{R^*}$ and $\bar{x}_s^*$. Thus, another estimator of $I_s^{R_p}$ is given by

$$\hat{I}_s^{A_p^*} = \frac{\bar{p}_s^{A^*}}{\bar{x}_s^* \hat{\beta}_0^A} \qquad \text{and} \qquad \hat{I}_s^{G_p^*} = \frac{\bar{p}_s^{G^*}}{\exp\left(\bar{x}_s^* \hat{\beta}_0^G\right)}, \tag{6}$$

which, relative to other hedonic estimators, presents the advantage of simplifying and reducing the costs of data collection and increasing the robustness and precision. The last two advantages are illustrated in the simulation study of the next section.

3 Monte Carlo illustration of robustness and efficiency gains

This section presents some Monte Carlo simulation experiments involving two geometric QAPI estimators: $\hat{I}_s^{G_p^*}$ of (6), that uses aggregate population information on both prices and characteristics at the index period s, and $\hat{I}_s^{G_p}$ of (5), that uses only one data source for all periods.

3.1 Experimental designs

Asset prices and characteristics are simulated for two periods, 0 and 1. For each period, the following loglinear hedonic function is used to generate asset prices: $\ln p_{it} = \beta_{t,0}^G + X_{it,1}\beta_{t,1}^G + X_{it,2}\beta_{t,2}^G + X_{it,3}\beta_{t,3}^G + u_{it}^G$, where $(X_{t,1}, X_{t,2})$ follow a multivariate normal distribution with means $(4.5, 5.0)$ and $(4.52, 5.01)$ at periods 0 and 1, respectively, and variances of 0.5 and null covariances at both periods; $X_{t,3}$ is a dummy variable that takes the value 1 with a probability of 0.38 $(t = 0)$ and 0.40 $(t = 1)$; and u_t^G is generated from a normal distribution with mean 0 and variance 1 at both periods. We set $\beta_0^G = \left(1, \beta_{0,1}^G, 1, 1\right)$ and $\beta_1^G = \left(1.00846, \beta_{0,1}^G + 0.015, 1.005, 0.985\right)$, which implies that $I_1^{G_p} = 1.1$. Unless otherwise stated below, $\beta_{0,1}^G = 1$ and $N_0 = N_1 = 1000$. All simulation results are based on 100 000 Monte Carlo replications.

To compute the two alternative QAPI estimators, we make the following assumptions. In terms of macro data, we assume knowledge on the population means of asset prices and characteristics for period 1. Hence, for calculating $\hat{I}_1^{G_p^*}$ we consider $\bar{x}_1^* = (1, 4.52, 5.01, 0.40)$ and $\bar{p}_s^{G^*} = \exp\left(\bar{x}_1^* \beta_1^G\right)$. In terms of micro data, we assume the availability of a dataset for period 0 that allows consistent estimation of the parameters β_0^G, while for period 1 we consider three distinct sets of experiments:

Experiment 1: *Absence of sampling problems*

In the first experiment it is assumed that there are no sampling problems and that a dataset that would also allow consistent estimation of β_1^G is available. This experiment is used both to illustrate the potential gains of precision that the use of macro information may originate and to act as a benchmark for some of the remaining experiments. Three different sample sizes are considered: $N_0 = N_1 = \{50, 100, 1000\}$.

Experiments 2-3: *Measurement error*

Measurement error may affect both the asset price and characteristics and, thus, may cause the inconsistency of $\hat{I}_1^{G_p}$. In both cases, we assume that instead of z_{i1}, the available micro sample contains information on $\dot{z}_{i1} = z_{i1} - e_{i1}$, where $z_{i1} = \ln p_{i1}$ (Experiment 2) or $z_{i1} = X_{it,1}$ (Experiment 3) and e_{i1} is the unobservable measurement error with mean μ_e and variance σ_e^2. We set $e_{i1} = \sigma_e\left(\xi_{i1} - 1\right) + \mu_e$, where ξ_{i1} was generated as an exponential variate with mean and variance one and $\left(\mu_e, \sigma_e^2\right) = \{(0, 1), (0, 2), (0.1, 1), (0.2, 1)\}$.

Experiments 4-5: *Missing data*

Two patterns of missing data are considered to illustrate the effects over the consistency of $\hat{I}_1^{G_p}$ of two (naive) strategies commonly used in applied work to deal with that problem: a strategy that discards all assets with missing values (Experiment 4 - this is a common procedure in cases where the missing values affect only some assets) and a strategy that discards all variables displaying missing values (Experiment 5 - this is a common procedure in cases where the missing values affect only some specific variables).

In Experiment 4 we divide the data in two subsamples, one containing the least expensive assets (subsample A) and the other the remaining assets (subsample B). Define $P_A = \Pr[r_i = 1 | p_{i1} \leq \text{median}(p_{i1})]$ and $P_B = [r_i = 1 | p_{i1} > \text{median}(p_{i1})]$, where r is an indicator variable that takes the value 1 if no asset information is missing. After randomly generating a sample of N_1 assets, we drew random samples of sizes $P_A N_1/2$ and $P_B N_1/2$ from subsamples A and B, respectively, in order to form a sample of $(P_A + P_B) N_1/2$ observations, with the remaining observations being discarded. We consider $(P_A, P_B) = \{(0.5, 0.5), (0.25, 0.25), (0.5, 0.6), (0.5, 0.7)\}$.

Concerning Experiment 5, we assume that only $X_{i1,1}$ has missing values and that the analyst decides to omit from the formula defining $\hat{I}_1^{G_p}$ not only $X_{i1,1}$ but also $X_{i0,1}$, reestimating β_0^G using only the remaining covariates. The design parameter is $\beta_{0,1}^G = \{0, 2.5, 5\}$.

3.2 Results

Table 1 displays the mean and the standard deviation across replications of both $\hat{I}_1^{G_p^*}$ and $\hat{I}_1^{G_p}$. All results illustrate clearly the benefits of using macro data whenever available, both in terms of robustness and precision. Even in the absence of sampling problems, the precision gains may achieve 30%. In this case, using macro information has also the advantage of attenuating the small bias displayed by the standard estimator $\hat{I}_1^{G_p}$ for the sample sizes of 50 and 100.

Under sampling problems that affect only the micro data set collected for period 1, the performance of $\hat{I}_1^{G_p^*}$ obviously does not change at all, while $\hat{I}_1^{G_p}$ may or may not become inconsistent, depending on the particular sampling problem simulated. In particular, any sampling problem that does not change the mean of both the asset prices and asset characteristics leaves the consistency of $\hat{I}_1^{G_p}$ unaffected, as could be anticipated from (5). This is the case of additive measurement error with mean zero (two first examples of Experiments 2 and 3), data missing-completely-at-random (two first examples of Experiment 4) and omission of an irrelevant covariate (first example of Experiment 5). However, even in these cases the efficiency gains of exploiting macro information range from 28% to 51%, since large measurement error variances and large amounts of missing data decrease substantially the precision of the analysis.

In all cases where the sample mean of the asset prices and/or asset characteristics is an inconsistent estimator of the corresponding means in the population, $\hat{I}_1^{G_p}$ is also an inconsistent estimator of the QAPI. Naturally, larger deviations of the mean of the measurement error from zero, larger distortions between the sample and the population structures caused by missing data and larger contributions of the omitted variable to the asset price lead to higher bias in the estimation of QAPI.

References

Collins, A., Scorcu, A. and Zanola, R. (2009), "Reconsidering hedonic art price indexes", *Economics Letters*, 104(2), 57-60.

Georges, P. and Seçkin, A. (2013), "Black notes and white noise: a hedonic approach to auction prices of classical music manuscripts", *Journal of Cultural Economics*, 37(1), 33-60.

Hill, R.J. and Melser, D. (2008), "Hedonic imputation and the price index problem: an application to housing", *Economic Inquiry*, 46(4), 593-609.

Table 1: Monte Carlo QAPI estimates

		Mean		St.Dev.	
		$\hat{I}_1^{P*}$	$\hat{I}_1^{P}$	$\hat{I}_1^{P*}$	$\hat{I}_1^{P}$
Experiment 1: Absence of sampling problems					
$N =$	50	1.111	1.123	0.164	0.236
	100	1.106	1.111	0.113	0.161
	1000	1.101	1.101	0.035	0.049
Experiment 2: Price measurement error					
$(\mu_e, \sigma_e) =$	$(0, 1)$	1.101	1.102	0.035	0.061
	$(0, 2)$	1.101	1.102	0.035	0.070
	$(0.1, 1)$	1.101	0.997	0.035	0.055
	$(0.2, 1)$	1.101	0.902	0.035	0.050
Experiment 3: Covariate measurement error					
$(\mu_e, \sigma_e) =$	$(0, 1)$	1.101	1.102	0.035	0.061
	$(0, 2)$	1.101	1.102	0.035	0.070
	$(0.1, 1)$	1.101	1.218	0.035	0.067
	$(0.2, 1)$	1.101	1.346	0.035	0.075
Experiment 4: Missing obvervations					
$(P_A, P_B) =$	$(0.5, 0.5)$	1.101	1.102	0.035	0.058
	$(0.25, 0.25)$	1.101	1.102	0.035	0.071
	$(0.5, 0.6)$	1.101	1.156	0.035	0.059
	$(0.5, 0.7)$	1.101	1.204	0.035	0.060
Experiment 5: Missing covariates					
$\beta_{0,1}^G =$	0	1.101	1.101	0.035	0.049
	2.5	1.101	1.161	0.035	0.106
	5	1.101	1.231	0.035	0.204

Imbens, G.W. and Lancaster, T. (1994), "Combining micro and macro data in microeconometric models", *Review of Economic Studies*, 61, 655-680.

Pakes, A. (2003), "A reconsideration of hedonic price indexes with an application to PC's", *American Economic Review*, 93(5), 1578–1596.

Ramalho, E.A. and Ramalho, J.J.S. (2014), "Convenient links for the estimation of hedonic price indexes", *Statistica Neerlandica*, 68(2), 91-117.

Reis, H. and Santos Silva, J.M.C. (2006), "Hedonic indexes for new passenger cars in Portugal (1997-2001)", *Economic Modelling*, 23(6), 890-908.