


Combining micro and macro data in hedonic price indexes

Esmeralda A. Ramalho¹ · Joaquim J. S. Ramalho¹  · Rui Evangelista²

Accepted: 22 August 2016 / Published online: 30 August 2016
© Springer-Verlag Berlin Heidelberg 2016

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 indexes have 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. A Monte Carlo simulation study and an empirical application concerning the housing market illustrate some of those advantages.

Keywords Paasche price index · Imputation hedonic method · Quality adjustment

JEL classification C43 · E31 · R31

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

The authors thank the Editor and the referee for their valuable suggestions and remarks that substantially improved the paper. The authors would like to also thank Statistics Portugal for providing the necessary data that made the empirical application presented in this paper possible. Financial support from Fundacao para a Ciencia e a Tecnologia (Grants PTDC/EGE-ECO/119148/2010 and UID/ECO/04007/2013) is also gratefully acknowledged.

✉ Joaquim J. S. Ramalho
jsr@uevora.pt

¹ Department of Economics and CEFAGE, Universidade de Évora, Évora, Portugal

² Statistics Portugal, Lisboa, Portugal

(see, e.g., Hill and Melser 2008), artworks (e.g., Collins et al. 2009) and collectables (e.g., Georges and Seçkin 2013). 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 are often 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.

This paper is organized as follows. Section 2 discusses how aggregate data may be used to construct Paasche QAPI. Section 3 presents a Monte Carlo illustration of the efficiency and robustness gains of the proposed index. An empirical application concerning the housing market is provided in Sect. 4. Section 5 concludes.

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 , $j = 1, \dots, k$, and let x_{it} be the $1 \times (k + 1)$ vector with elements $X_{it,j}$, $j = 0, \dots, k$, 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:

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

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 \(2014\)](#) for a comprehensive analysis of this issue. 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

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

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:

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

$$\hat{I}_s^{A_q} = \frac{\frac{1}{N_s} \sum_{i=1}^{N_s} x_{is} \hat{\beta}_s^A}{\frac{1}{N_0} \sum_{i=1}^{N_0} x_{i0} \hat{\beta}_0^A}, \quad \hat{I}_s^{G_q} = \frac{\exp \left(\frac{1}{N_s} \sum_{i=1}^{N_s} x_{is} \hat{\beta}_s^G \right)}{\exp \left(\frac{1}{N_0} \sum_{i=1}^{N_0} x_{i0} \hat{\beta}_0^G \right)} \quad (3)$$

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 eq. 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 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 $\sum_{i=1}^{N_t} p_{it} = \sum_{i=1}^{N_t} \hat{p}_{it} = \sum_{i=1}^{N_s} x_{it} \hat{\beta}_t^A$ (arithmetic indexes) and $\sum_{i=1}^{N_t} \ln(p_{it}) = \sum_{i=1}^{N_t} \ln(\hat{p}_{it}) = \sum_{i=1}^{N_s} x_{it} \hat{\beta}_t^G$ (geometric indexes). That is, the price averages of the current period are numerically equal to the average of the product of characteristics and shadow prices of the current period. Hence, (2) may be simplified to:

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

where \bar{p}_s^R denotes the sample arithmetic or geometric mean of the asset prices in the current period. Hence, unlike suggested by eq. (2), the hedonic function needs to be estimated only at the base period. Indeed, because of the numerical equivalence between indexes (4) and (2), the shadow prices of the asset characteristics are still implicitly allowed to change over time in (4), although we do not need to estimate them.

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:

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

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).

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.¹ 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 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

$$\hat{I}_s^{A_p} = \frac{\bar{p}_s^{A*}}{\bar{x}_s^* \hat{\beta}_0^A} \quad \text{and} \quad \hat{I}_s^{G_p} = \frac{\bar{p}_s^{G*}}{\exp(\bar{x}_s^* \hat{\beta}_0^G)}. \quad (6)$$

The possibility of combining information from different sources in the index formulas (6) presents several advantages. On the one hand, the effort on data collection and organization at period s is substantially reduced with the use of aggregate information. A first issue dealt with is the reluctance of several data providers in releasing individual data due to confidentiality issues. In fact, due to the sensitiveness of some kinds of individual information, micro data release typically involves an investment in disclosure protection treatments, which must be carefully applied in order to avoid the introduction of bias and increased variability in secondary statistical analysis; see [Reiter \(2005, 2012\)](#) for a detailed discussion. Another important difficulty in data preparation for the construction of hedonic price indexes is the combination of individual information from different sources to obtain all relevant variables.² While the inclusion of aggregate information in the index formulas (6) is straightforward, the matching processes for the combination of micro datasets at each period s are typically complex, being considered error-prone and producing often samples with important amounts of missing data due to unmatched observations; see e.g. [Ridder and Moffit \(2007\)](#).

On the other hand, robustness and efficiency gains may arise from the use of aggregate information on the population of interest or from large samples not available in an individual basis. This idea builds on the general literature on the combination of macro information on micro models, pioneered by [Imbens and Lancaster \(1994\)](#). These authors show how to incorporate information on aggregate quantities of the population of interest, assumed to have little or no sampling error, in the estimation of moment condition models based on individual data. The macro information is added to the estimation procedure through overidentifying restrictions, producing efficiency gains and reinforcing the robustness of microeconomic estimators to problems such as variable omission, measurement error and endogenous sampling; see, respectively, [Hellerstein and Imbens \(1999\)](#), [Ramalho \(2002\)](#) and [Ramalho and Ramalho \(2006\)](#). In this paper, the approach is different: the aggregate information is not considered for

¹ Note that only the means of the characteristics that are relevant at the base period are necessary.

² For example, Statistics Finland uses micro data from three different sources to construct house price indexes; see [Saarnio \(2006\)](#).

the estimation of the hedonic function at the base period but incorporated directly in the index formulas (6) for all periods $t \neq 0$. Robustness and precision gains may arise in this context relative to the indexes (5) because of the use of exact, or nearly-exact, aggregate information, which may avoid various sampling issues that typically affect estimators based on micro data.³

3 Monte Carlo illustration of robustness and efficiency gains

This section presents some Monte Carlo simulation experiments that illustrate how the inclusion of aggregate population information in the QAPI formula may produce gains of precision and/or robustness relative to estimators based exclusively on micro data. In particular, the geometric QAPI estimator that uses aggregate information on both prices and characteristics at the index period s , $\hat{I}_s^{G^*}$ of (6), is compared with $\hat{I}_s^{G^p}$ 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 log-linear hedonic function is used to generate asset prices:

$$\ln p_{it} = \beta_{t,0}^G + X_{it,1}\beta_{t,1}^G + X_{it,2}\beta_{t,2}^G + X_{it,3}\beta_{t,3}^G + u_{it}^G, \quad (7)$$

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 $\beta_0^G = (1, \beta_{0,1}^G, 1, 1)$ and $\beta_1^G = (1.00846, \beta_{0,1}^G + 0.015, 1.005, 0.985)$, 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 $\hat{I}_1^{G^*}$ we consider $\bar{x}_1^* = (1, 4.52, 5.01, 0.40)$ and $\bar{p}_s^{G^*} = \exp(\bar{x}_1^* \beta_1^G)$. 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:

³ Note that the incorporation of additional restrictions *also* in the moment condition version of the OLS estimator at moment 0, following directly [Imbens and Lancaster \(1994\)](#), would potentially yield more robust and precise estimators for β_0 . However, eqs. (4) and (5) would no longer hold and, hence, the computation of QAPI would require direct application of the original formula (2) and the consequent estimation of one regression at each index period.

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 is an unfortunate feature of micro data that may affect both the asset price and characteristics and, thus, may cause the inconsistency of \hat{I}_1^{Gp} . In both cases, we assume that instead of z_{i1} , the available micro sample contains information on $\tilde{z}_{i1} = z_{i1} - e_{i1}$, 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 $e_{i1} = \sigma_e (\xi_{i1} - 1) + \mu_e$, where ξ_{i1} was generated as an exponential variate with mean and variance one and $(\mu_e, \sigma_e^2) = \{(0, 1), (0, 2), (0.1, 1), (0.2, 1)\}$.

Experiments 4–5: Missing data

The existence of missing values in asset prices and/or asset characteristics is endemic; see *inter alia* the application to the housing market of Hill and Syed (2016). This third set of experiments considers two examples of missing data, illustrating the effects over the consistency of \hat{I}_1^{Gp} 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^{Gp} 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*} and \hat{I}_1^{Gp} . 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 informa-

Table 1 Monte Carlo QAPI estimates

		Mean		SD	
		$\hat{I}_1^{G_p^*}$	$\hat{I}_1^{G_p}$	$\hat{I}_1^{G_p^*}$	$\hat{I}_1^{G_p}$
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 observations					
$(P_A, P_B) =$	(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

tion 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.

4 Empirical application: price indexes for apartments in Lisbon, Portugal

This section illustrates the use of the approach proposed in this paper to produce a Paasche QAPI for apartments in the municipality of Lisbon, Portugal. First, we describe the data provided by Statistics Portugal for this application, which is a subset of the data that are currently being used by that institute to construct the official house price index for Portugal. Then, we show how the process of constructing a QAPI for the Lisbon housing market may be carried out in a simplified way.

4.1 Data

At the moment, Statistics Portugal uses two different data sources to compile the official house price index for Portugal (INE 2014). Both databases consist of administrative records generated for property transfer and local property taxes purposes and are maintained by the Portuguese Tax and Customs Authority. The first data source contains information that is relevant for the calculation of the Municipal Tax on Real Estate Transfers (IMT), which is a tax levied on property transfers. The IMT is calculated based on the value of the transaction (declared in the sales deed) or on the updated fiscal appraisal value of the property, depending on which is higher. This system implies that the recorded transaction values are the same or close to real transaction values. Because it represents a non-negligible cost to the buyer, IMT is typically paid just a few days before or on the same day the property is transacted. Therefore, the date of IMT payment constitutes a trustworthy indicator of the transaction moment. Moreover, the whole population of transactions is covered, since a proof of the payment of the IMT has to be shown by the buyer before a sale takes place. On the other hand, this database does not include dwelling characteristics. Therefore, apart from information characterizing the type and purpose of the transaction, the information received by Statistics Portugal from this database includes only the transaction price, the transaction moment and, which is crucial for the matching processes undertaken by Statistics Portugal, the property cadastral register identification number.

The second data source consists of the Local Property Tax (IMI) records. The IMI is a municipal tax levied on the current appraisal value of the dwelling, which is computed from a formula given in the Portuguese Property Tax Code that is a function of the dwelling characteristics recorded in the IMI database. Because each dwelling is identified by the same property cadastral register identification number used in IMT records, it is possible to match the two sources of data to produce a unique dataset containing both house prices and characteristics. It is this unique dataset that is used by Statistics Portugal in its index compilations. Any unmatched transaction is excluded from the index computation.

Currently, Statistics Portugal applies the adjacent-period time dummy method to produce a unique house QAPI for Portugal, not producing any regional indexes. In contrast, in this paper we apply the imputation price method, which does not require the assumption of parameter constancy over two or more time periods, and consider only apartments and the municipality of Lisbon, which displays the highest number of residential transactions in Portugal. We use quarterly data provided by Statistics Portugal for the period 2009–2013. During this 5-year period, Lisbon accounted for around 7 % of the total transactions and 9 % of all apartment sales that took place in Portugal. The data supplied by Statistics Portugal includes all IMT data of apartment transactions carried out in Lisbon and some of the dwelling characteristics contained in the matched IMI records. The supplied IMT database provides information on 32,156 apartment transactions occurred from 2009 to 2013 in Lisbon, from which 27,958 (86.9 %) were matched with the IMI data.

There are three main reasons that explain the existence of unmatched transactions. The first one has to do with the fact that, although covering almost entirely the stock of residential properties, the IMI data made available to Statistics Portugal by the tax authority does not cover the appraisals carried out from December 2003 (i.e., when the tax was first introduced) to December 2004. As a result of this, there may exist properties left unmatched simply because the IMI information on its characteristics was collected during that period for which it was impossible to obtain records from the tax authority. The second reason stems from the nature of the IMI, which generates information that is continuously subject to update because, among other reasons, it can be contested by tax payers. As mentioned above, the payment of the IMT involves the comparison of transaction and appraised property values. If tax payers do not agree with an appraised value, it is possible to ask for a revaluation and the transaction stays with no paired IMI information until the appraisal is considered as final. The IMI information can also be contested or unavailable for other reasons, such as when taxpayers find mistakes in appraisals (e.g., in the age of the property). The time taken to solve these issues vary. As a consequence, since the IMI data used in this paper refers to an extraction done at a particular point in time (i.e., 2014), it is natural to see a tendency to have a growing percentage of unmatched transactions as one moves towards the end of 2013 due to these reasons. Finally, some transactions may be left unmatched due to the existence of errors in property identification numbers or because there is no correspondence between the end use of the property in IMT and IMI records (which could also be caused by mistakes in declared end uses).

In order to examine some consequences of the matching process, Table 2 presents: (i) the distribution of IMT (N_s^*) and matched IMT/IMI (N_s) data across the 16 quarters in analysis; (ii) the arithmetic average of the transaction prices for both IMT (\bar{p}_s^*) and IMT/IMI (\bar{p}_s); and (iii) the unadjusted arithmetic price indexes based on each database (\hat{I}_s^* and \hat{I}_s), which were calculated according to (1), with 2009Q1 being the base quarter. Percentage differences between the matched and the IMT databases are also reported for all quantities.

From Table 2, it is apparent that the reduction in the sample size due to the matching process becomes more relevant in the second half of the period in analysis. On the other hand, the average annual transaction price in the matched IMT/IMI data is

Table 2 Consequences of the matching process

Quarter	N_s^*	\bar{P}_s^*	\hat{I}_s^*	N_s	\bar{P}_s	\hat{I}_s	Difference (%)	
							N_s vs N_s^*	\bar{P}_s vs \bar{P}_s^*
							\hat{I}_s vs \hat{I}_s^*	
							N_s vs N_s^*	\bar{P}_s vs \bar{P}_s^*
2009Q1	1475	193,017.5	1.000	1313	196,852.4	1.000	-11.0	2.0
2009Q2	1695	202,358.6	1.048	1524	204,972.5	1.041	-10.1	1.3
2009Q3	2075	202,699.3	1.050	1870	202,659.3	1.029	-9.9	0.0
2009Q4	2129	194,968.8	1.010	1887	199,730.8	1.015	-11.4	2.4
2010Q1	2260	209,539.4	1.086	2025	216,517.8	1.100	-10.4	3.3
2010Q2	2178	196,918.8	1.020	1926	205,601.5	1.044	-11.6	4.4
2010Q3	1893	200,377.7	1.038	1702	205,205.8	1.042	-10.1	2.4
2010Q4	2263	187,637.3	0.972	2053	191,370.9	0.972	-9.3	2.0
2011Q1	1693	193,230.4	1.001	1494	200,980.6	1.021	-11.8	4.0
2011Q2	1670	223,606.4	1.158	1469	231,844.7	1.178	-12.0	3.7
2011Q3	1175	197,608.6	1.024	1012	198,204.8	1.007	-13.9	0.3
2011Q4	1367	203,283.7	1.053	1147	201,591.8	1.024	-16.1	0.8
2012Q1	1171	208,702.8	1.081	975	223,686.2	1.136	-16.7	7.2
2012Q2	1205	189,789.1	0.983	976	193,854.1	0.985	-19.0	2.1
2012Q3	961	183,875.1	0.953	790	189,703.3	0.964	-17.8	3.2
2012Q4	1253	191,398.6	0.992	1007	199,917.9	1.016	-19.6	4.5
2013Q1	1082	193,175.3	1.001	894	206,070.5	1.047	-17.4	6.7
2013Q2	1297	204,851.4	1.061	1077	220,061.4	1.118	-17.0	7.4
2013Q3	1469	213,017.3	1.104	1245	223,779.4	1.137	-15.2	5.1
2013Q4	1845	208,831.4	1.082	1572	219,429.9	1.115	-14.8	5.1
Total	32,156	200,366.2		27,958	206,710.3		-13.1	3.2

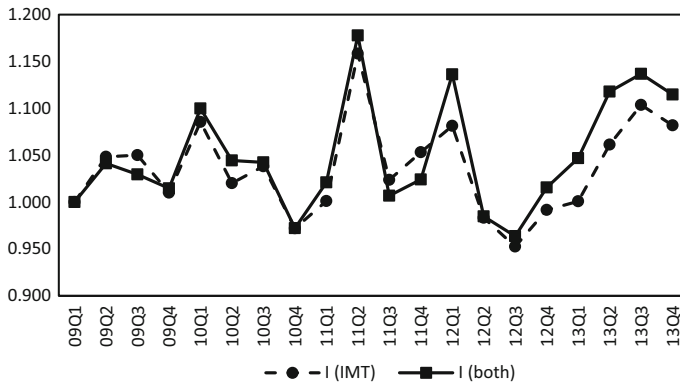


Fig. 1 Unadjusted house price indexes

systematically larger than that of IMT, with an average difference of 3.2 % over the whole 5-year period. This difference becomes clearly more important in the last 2 years of the sample, where the average difference is 5.1 %, which more than doubles the corresponding figure for the first 3 years (2.1 %). Thus, while the matching process appears to generate a sample that is approximately random in the first quarters, the same does not seem to occur for the remaining period. This conjecture is corroborated by the unadjusted arithmetic price indexes obtained from each database. Indeed, the differences between both indexes are relatively small before 2012, but afterwards the IMT/IMI index appears to be clearly inflated relative to the one based solely on IMT; see also Fig. 1. This suggests that an index of the type proposed in this paper may be useful in this framework, given that, for house prices, the matching process is only required for a single period and in the others the aggregate price information from IMT may be used. However, note that for an ideal application of our approach we would also need aggregate information on dwelling characteristics, which is not available from the IMT database. Therefore, the aggregate index constructed in the next section still requires the matching data for all periods in which concerns the dwelling attributes.

4.2 Index construction

While the simulation study of Sect. 3 focussed on geometric indexes, now we illustrate the construction of indexes based on arithmetic means. Actually, in practice, it will be much more common to have immediate access to the latter type of average than the former. We consider two alternative Paasche QAPI. The first (\hat{I}_s^{Ap}) is based exclusively on the matched IMT/IMI data, being calculated as in (2), where an hedonic function needs to be estimated for each quarter, or as in (4) or (5), where the hedonic function is estimated only at the base quarter (2009Q1). The three formulas produce numerically identical results. The second index (\hat{I}_s^{A+}) is based on a modified version of (6), since we do not have IMT data on dwelling characteristics:

$$\hat{I}_s^{A+} = \frac{\bar{p}_s^*}{\bar{x}_s \hat{\beta}_0^A}. \quad (8)$$

Thus, in this second case we use the IMT/IMI data to obtain averages of dwelling characteristics and IMT data for price averages. For the base quarter we need to estimate the hedonic function using the same micro data considered for the first index. The advantage of $\hat{I}_s^{A_p^+}$ over $\hat{I}_s^{A_p}$ is that it relies on more accurate measures of apartment prices, since IMT covers the whole population of interest.

Unlike what has been implicitly assumed throughout this paper, the matching in the base period is not perfect. Therefore, the prices predicted by the hedonic function, $\bar{x}_0 \hat{\beta}_0^A$, equal the IMT/IMI price average \bar{p}_0 but not the IMT price average \bar{p}_0^* , which implies that $\hat{I}_0^{A_p^+} \neq 1$. Hence, we applied the following rescaling:

$$\hat{I}_s^{A_p^{r+}} = \frac{\hat{I}_s^{A_p^+}}{\hat{I}_0^{A_p^+}}, \quad (9)$$

which ensures that $\hat{I}_s^{A_p^{r+}} = \hat{I}_s^{A_p} = 1$. Note that the same quarterly house price changes are produced by both $\hat{I}_s^{A_p^+}$ and $\hat{I}_s^{A_p^{r+}}$. Working with the rescaled version (9) is just a matter of convenience, since it allows direct comparisons between $\hat{I}_s^{A_p}$ and $\hat{I}_s^{A_p^{r+}}$.

The estimated linear hedonic model for 2009Q1 is the following:

$$\begin{aligned} \hat{p}_{i2009Q1} = & -59971.0 + 2091.5GRAREA_i + 2129.9DEPAREA_i - 788.4AGE_i \\ & (17832.0) \quad (146.8) \quad (293.6) \quad (207.1) \\ & + 5.0AGE_{it}^{0.5} + 21393.0DWATERF_i + 107199.0DSCENIC_i \\ & (1.7) \quad (6008.1) \quad (16989.0) \\ & + 26771.0DEXCPLOC_i, R^2 = 0.698, N = 1313 \\ & (4226.0) \end{aligned} \quad (10)$$

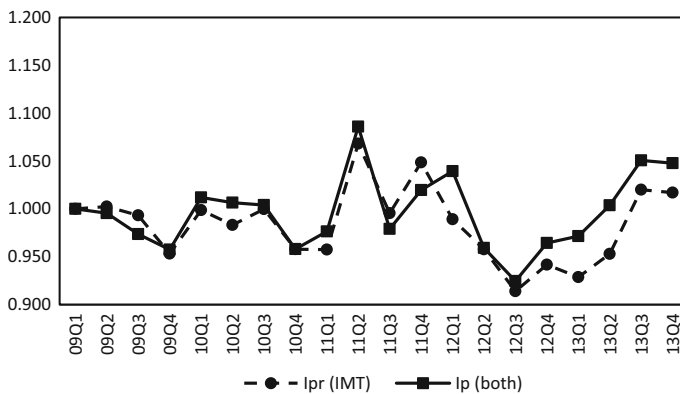
where *GRAREA* is the gross floor area of an apartment, *DEPAREA* provides the total area of its dependent areas (e.g., garages, cellars), *AGE* is the number of years of the apartment at transaction date and *DWATERF*, *DSCENIC* and *DEXCPLOC* are 0/1 variables signaling apartments located, respectively, in a Lisbon parish with access to a waterfront area, in places with scenic or visual value and in zones with an extremely good offer in terms of accessibilities, public transport and other infrastructures and amenities; see Table 3 for descriptive statistics of the dwelling characteristics.

Table 3 Dwelling attribute description

Variable	Unit	Mean	SE
<i>GRAREA</i>	m^2	103.042	51.474
<i>DEPAREA</i>	m^2	16.554	20.045
<i>AGE</i>	#	32.653	32.669
<i>DWATERF</i>	1/0	0.391	0.488
<i>DSCENIC</i>	1/0	0.004	0.061
<i>DEXCPLOC</i>	1/0	0.842	0.365

Table 4 Paasche QAPI

Quarter	$\hat{I}_s^{A_p}$	$\hat{I}_s^{A_p^{r+}}$
2009Q1	1.000	1.000
2009Q2	0.996	1.002
2009Q3	0.974	0.993
2009Q4	0.957	0.953
2010Q1	1.012	0.999
2010Q2	1.007	0.983
2010Q3	1.004	1.000
2010Q4	0.958	0.958
2011Q1	0.976	0.957
2011Q2	1.086	1.068
2011Q3	0.979	0.996
2011Q4	1.020	1.049
2012Q1	1.039	0.989
2012Q2	0.959	0.958
2012Q3	0.925	0.914
2012Q4	0.964	0.952
2013Q1	0.971	0.929
2013Q2	1.004	0.953
2013Q3	1.051	1.020
2013Q4	1.048	1.017

**Fig. 2** Quality-adjusted house price indexes

All explanatory variables are individually significant to explain the transaction price at the 1 % level and their coefficients have the expected sign.

Using the regression coefficients of (10) and quarterly averages of prices and characteristics, we obtained the $\hat{I}_s^{A_p}$ and $\hat{I}_s^{A_p^{r+}}$ indexes reported in Table 4 and Fig. 2. As expected, the price level appears overestimated by the matched IMT/IMI, $\hat{I}_s^{A_p}$, which

presents a price increase relative to the base quarter in nine out of twenty quarters. In contrast, the proposed QAPI $\hat{I}_s^{A^+}$ only shows that trend in five quarters and reveals apartment average prices above base prices for two consecutive quarters only in the last half of 2013.

5 Conclusion

In this paper we proposed a new approach for constructing QAPI for infrequently traded heterogeneous assets that has several advantages over alternative hedonic methods in terms of data requirements, robustness and precision. 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 $\hat{I}_s^{R_q} = \exp(\bar{x}_s \hat{\beta}_0) / \bar{p}_0^R$. In this case, it is not even required any type of information on asset prices for the index period. Similarly, while in this paper we focused on the imputation price method, our approach can also be applied to re-pricing hedonic indexes, which, for example, are used in the compilation of the official house price index in Slovenia; see Pavlin (2015). In fact, because re-pricing indexes result from the ratio of unadjusted price indexes and quality indexes based on hedonic coefficients of the base period, it is clear that the only micro information required by them is the one employed in the estimation of the hedonic model at the base period.

Another possible extension of our approach is its application to the measurement of productivity or other differences *across groups* (e.g., regions, sectors of activity, gender). This is specially relevant for the strand of literature that decomposes the overall mean differences of logged outcomes of two groups into a component that reflects differences in the observable group characteristics and another component that is attributed to other causes (e.g., discrimination); see the recent survey paper of Fortin et al. (2011). Clearly, this decomposition is akin to that analyzed in this paper, with 0 representing the reference group and s indexing other group. Several important refinements have been proposed in this area, namely the extension of the decomposition for distributional parameters other than the mean, but none of those refinements restricts the estimation and micro data requirements to the reference group.

References

- Collins A, Scorcu A, Zanola R (2009) Reconsidering hedonic art price indexes. *Econ Lett* 104(2):57–60
- Fortin N, Lemieux T, Firpo S (2011) Decomposition methods. In: Ashenfelter O, Card D (eds) *Handbook of labor economics*, vol 4A. Elsevier, Amsterdam, pp 1–102
- Georges P, Seçkin A (2013) Black notes and white noise: a hedonic approach to auction prices of classical music manuscripts. *J Cultural Econ* 37(1):33–60
- Hellerstein JK, Imbens GW (1999) Imposing moment restrictions from auxiliary data by weighting. *Rev EconStat* 81:1–14
- Hill RJ, Melser D (2008) Hedonic imputation and the price index problem: an application to housing. *Econ Inq* 46(4):593–609
- Hill RJ, Syed IA (2016) Hedonic price-to-rent ratios, user cost, and the detection of departures from equilibrium in the housing market. *Reg Sci Urb Econ* 56:60–72

- Imbens GW, Lancaster T (1994) Combining micro and macro data in microeconomic models. *Rev Econ Stud* 61:655–680
- INE - Instituto Nacional de Estatística (2014), Índice de Preços da Habitação: Documento Metodológico (Versão 1.0). Available only in Portuguese. <http://smi.ine.pt/DocumentacaoMetodologica/Detalhes/1269>
- Pakes A (2003) A reconsideration of hedonic price indexes with an application to PC's. *Am Econ Rev* 93(5):1578–1596
- Pavlin B (2015) House price indexes, Slovenia. <http://www.stat.si/StatWeb/Common/PrikaziDokument.ashx?IdDatoteke=8346>
- Ramalho EA (2002) Regression models for choice-based samples with misclassification in the response variable. *J Econom* 106(1):171–201
- Ramalho EA, Ramalho JJS (2006) Two-step empirical likelihood estimation under stratified sampling when aggregate information is available. *Manch Sch* 74(5):577–592
- Ramalho EA, Ramalho JJS (2014) Convenient links for the estimation of hedonic price indexes. *Stat Neerl* 68(2):91–117
- Reis H, Santos Silva JMC (2006) Hedonic indexes for new passenger cars in Portugal (1997–2001). *Econ Modell* 23(6):890–908
- Reiter JP (2005) Estimating risks of identification disclosure for microdata. *J Am Stat Assoc* 100:1103–1113
- Reiter JP (2012) Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public Opin Q* 76:163–181
- Ridder G, Moffit R (2007) The econometrics of data combination. In: Heckman JJ, Leamer EE (eds) *Handbook of econometrics*, vol 6B. Elsevier, Amsterdam, pp 5469–5547
- Saarnio M (2006) Housing price statistics at statistics Finland, Paper presented at the OECD-IMF workshop on real estate price indexes, Paris 6–7 November 2006