Statistica Neerlandica (2014) Vol. 68, nr. 2, pp. 91–117
doi:10.1111/stan.12024

Convenient links for the estimation of hedonic price indexes: the case of unique, infrequently traded assets

Esmeralda A. Ramalho and Joaquim J.S. Ramalho*

Department of Economics, Universidade de Évora, Evora, Portugal and CEFAGE-UE, Universidade de Évora, Evora, Portugal

Hedonic methods are a prominent approach in the construction of quality-adjusted price indexes. This paper shows that the process of computing such indexes is substantially simplified if arithmetic (geometric) price indexes are computed based on exponential (log-linear) hedonic functions estimated by the Poisson pseudo-maximum likelihood (ordinary least squares) method. A Monte Carlo simulation study based on housing data illustrates the convenience of the links identified and the very attractive properties of the Poisson estimator in the hedonic framework.

Keywords and Phrases: hedonic price indexes, quality adjustment, retransformation, house prices, exponential regression, Poisson pseudo-maximum likelihood.

1 Introduction

As is well known, price indexes for infrequently traded heterogeneous assets such as houses, artworks, and collectable objects cannot be constructed simply by comparing the average price of the assets sold in each time period, as the result would be dependent on the particular mix of assets that happened to be sold in those periods. Instead, the heterogeneity of the assets has somehow to be taken into account in order to separate the influences of quality changes from pure price movements. One way to do this is using the so-called hedonic pricing methodology, which is the technique recommended for the housing market by the recent *Handbook on Residential Property Price Indices* (Eurostat, 2013, Ch. 12, p. 7) and has been widely applied not only to houses (e.g., Dorsey, Hu, Mayer and Wang, 2010; Hill, 2013; Hill and Melser, 2008; Malpezzi, Chun and Green, 1998) but also to artworks (e.g., Ashenfelter and Graddy, 2003; Campos and Barbosa, 2009; Chanel, Gérard-Varet and Ginsburg, 1996; Collins, Scorcu and Zanola, 2009) and collectables (e.g., wines – Fogarty and Jones, 2011; music manuscripts – Georges and Seçkin, 2013; and rare diamonds – Renneboog and Spaenjers, 2012).[1]

*jsr@uevora.pt

[1]See Eurostat (2013) for a survey on alternative methods for compiling quality-adjusted price indexes.

Published by Wiley Publishing, 9600 Garsington Road, Oxford OX4 2DQ, UK and 350 Main Street, Malden, MA 02148, USA.

Quality-adjusted price indexes (QAPIs) based on the hedonic methodology build upon the idea that different characteristics of an asset impact differently on its price. To measure those impacts, it is necessary to specify the so-called hedonic price function, which relates transaction prices to the relevant asset characteristics. Using regression techniques, it is then possible to estimate the implicit marginal prices of each asset characteristic. Finally, based on the estimated marginal prices, and using an appropriate method, asset prices can be adjusted in order to remove the effect of quality changes. Along this process, among other aspects, four important choices have to be made: (i) the type of price index to compute (e.g., geometric or arithmetic); (ii) the form of the hedonic function (e.g., price or logged price as the dependent variable; linear or non-linear specification for the right-hand side (RHS)); (iii) the method used to estimate the parameters of the hedonic function (e.g., ordinary or weighted least squares); and (iv) the hedonic method used to adjust the prices (e.g., imputation price method or time dummy variable method).

The choice of the functional form of the hedonic function, and its relationship with the choice of the price index, is one of the key issues in the general literature on hedonic price indexes. Most authors argue that those choices should be made in an independent way. Otherwise, would the former require a specific form for the latter, researchers could be forced to use a functional form that is inconsistent with the data, which might create an error in the quality adjustment procedure (see *inter alia* Triplett, 2006). In fact, as discussed by Pakes (2003) in a very interesting analysis of the hedonic approach, the hedonic equation is merely a reduced form equation determined by the interaction of supply and demand, and, hence, there are no *a priori* restrictions on its functional form: practitioners should simply use some statistical criteria to chose the hedonic function that provides the most accurate price predictions. In contrast, Reis and Santos Silva (2006) claimed that the type of price index that is used determines not only the form of the dependent variable of the hedonic regression but also the estimation method that should be used. Otherwise, some basic properties of QAPI would not be satisfied. Finally, there is an apparent consensus that the time dummy variable method should be used only in association with price indexes based on geometric means and hedonic functions where the logged price is the dependent variable and the RHS is linear in the parameters. This is because the main attractiveness of the time dummy variable method is the possibility of obtaining very simple expressions for QAPI, and *all* authors seem to think that such expressions can only be obtained using the mentioned combination of price index and hedonic function.

In order to clarify and conciliate some of these divergent positions, in this paper we investigate in a comprehensive way whether or not there exist any links between the type of price index to be computed and the form of hedonic functions, hedonic methods, and estimation methods. We consider that there is a link whenever a specific combination of price indexes, functions, and methods substantially simplifies the calculations required to compute hedonic price indexes. To simplify the analysis, we focus on two widely used hedonic price indexes, the arithmetic hedonic price index (AHPI) and the geometric hedonic price index (GHPI), and consider only unit value indexes, which are appropriate for the type of assets we are interested in (unique assets traded only once in the sampling period).[2] In contrast to the traditional practice in the hedonic literature, which focus exclusively on QAPI, throughout the paper we use a similar framework to that commonly applied in the analysis of the decomposition of mean outcome differences between groups (see, e.g., the recent survey by Fortin, Lemieux, and Firpo, 2011) and examine the decomposition of unadjusted price indexes into quality and quality-adjusted price components under a variety of assumptions.

Three main points emerge from our investigation: (i) there is a very convenient link between AHPI (GHPI) and hedonic functions that use the asset price (logged price) as dependent variable – failure in respecting this link implies that consistent estimation of QAPI requires in general the application of bias corrections; (ii) there is a link between a linear (exponential) specification for the RHS of the hedonic function and the ordinary least squares (OLS) (Poisson pseudo-maximum likelihood (PPML)) method – if other estimation methods are used, then inconsistencies between alternative forms of calculating QAPI may arise, and the process of producing these indexes may be more time consuming; and (iii) in the time dummy framework, there is a link between AHPI (GHPI), hedonic functions that use the asset price (logged price) as dependent variable, and linear (exponential) specifications for the RHS of the hedonic function – otherwise, the QAPI will not have a simple expression. Through a Monte Carlo simulation study, we illustrate both the very large biases that may arise in this framework if link (i) is not respected and the very attractive properties of the PPML estimator in the hedonic context.

This paper is organized as follows. Section 2 briefly reviews the construction of hedonic price indexes. Section 3 examines whether there exists any link between price index formulas and the form of the dependent variable in the hedonic function. Section 4 analyzes the existence of links involving the method chosen for estimating the hedonic function. Section 5 examines the specific case of the time dummy variable method. Section 6 is dedicated to the Monte Carlo investigation. Finally, Section 7 concludes.

2 The construction of hedonic price indexes: a brief overview

Throughout this paper, p_{it} denotes the price p of asset i at period t, where i indexes different assets in 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 the characteristic j of asset i at period t, $j=1,\ldots,k$, and let x_{it} be the

[2]Unit value indexes are also used in a more general hedonic framework. See Haan (2004) for an example.

$1 \times (k+1)$ vector with elements $X_{it,j}$, $j = 0, \ldots, k$, where variable $X_{it,0} = 1$ denotes the constant term of the hedonic regression. Next, we provide a brief overview of the construction of hedonic price indexes.

2.1 Arithmetic and geometric price indexes

The two main alternative elementary formulas for computing price indexes for unique assets traded only once in the sampling period are based on the ratio of (unweighted) arithmetic or geometric means of prices. Let I^A and I^G be, respectively, the *population* arithmetic and geometric price indexes, and let $\bar{I}^A$ and $\bar{I}^G$ be the corresponding *sample* estimators. For period s, the sample arithmetic price index is given by

$$\bar{I}_s^A = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s} p_{is}}{\frac{1}{N_0}\sum_{i=1}^{N_0} p_{i0}}, \tag{1}$$

while the sample geometric price index may be written as

$$\bar{I}_s^G = \frac{\prod_{i=1}^{N_s} p_{is}^{\frac{1}{N_s}}}{\prod_{i=1}^{N_0} p_{i0}^{\frac{1}{N_0}}} = \frac{\exp\left[\frac{1}{N_s}\sum_{i=1}^{N_s}\ln(p_{is})\right]}{\exp\left[\frac{1}{N_0}\sum_{i=1}^{N_0}\ln(p_{i0})\right]}. \tag{2}$$

It is straightforward to show that $\bar{I}_s^A$ is a consistent estimator of the population arithmetic index

$$I_s^A = \frac{E(p_s)}{E(p_0)}, \tag{3}$$

while $\bar{I}_s^G$ is a consistent estimator of the population geometric index

$$I_s^G = \frac{\exp\{E[\ln(p_s)]\}}{\exp\{E[\ln(p_0)]\}}. \tag{4}$$

The overall asset price change between periods 0 and s is due to the different characteristics of the assets sold in each period and/or pure price movements. Thus, both I_s^A and I_s^G may be decomposed into two components: a quality index ($I_s^{A_q}$ or $I_s^{G_q}$), which assumes that the implicit prices of the asset characteristics do not change over time and, hence, measures the price change that is due to changes in the asset characteristics; and a QAPI ($I_s^{A_p}$ or $I_s^{G_p}$), which assumes that the asset characteristics are constant across time and measures the price change that is due to changes in their prices. Thus, we may write the population arithmetic price index as $I_s^A = I_s^{A_q} \cdot I_s^{A_p}$ and the population geometric price index as $I_s^G = I_s^{G_q} \cdot I_s^{G_p}$, where

$$I_s^{A_q} = \frac{E[E(p_b|x_s)]}{E[E(p_b|x_0)]}, \qquad I_s^{A_p} = \frac{E[E(p_s|x_a)]}{E[E(p_0|x_a)]}, \tag{5}$$

$$I_s^{G_q} = \frac{\exp\langle E\{E[\ln(p_b)|x_s]\}\rangle}{\exp\langle E\{E[\ln(p_b)|x_0]\}\rangle}, \quad I_s^{G_p} = \frac{\exp\langle E\{E[\ln(p_s)|x_a]\}\rangle}{\exp\langle E\{E[\ln(p_0)|x_a]\}\rangle} \tag{6}$$

and $(a,b)=(0,s)$ or $(s,0)$. Note that when $(a,b)=(0,s)$ $[(a,b)=(s,0)]$, $I_s^{A_p}$ $\left[I_s^{G_p}\right]$ is a Laspeyres (Paasche) QAPI, as the comparison is based on the assets existing at the base (current) period.

The prices of the asset characteristics are not observable, so the sample estimators $\overline{I}_s^A$ and $\overline{I}_s^G$ cannot be directly decomposed into quality and QAPIs. However, if a sample of the assets characteristics is available for each period, it is possible to estimate their implicit prices, and their evolution, using the hedonic regression, which relates (asset) prices to (asset) characteristics. Based on this regression, alternative estimators for the unadjusted arithmetic and geometric price indexes may be constructed, being given by, respectively,

$$\hat{I}_s^A = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s}\hat{p}_{is}}{\frac{1}{N_0}\sum_{i=1}^{N_0}\hat{p}_{i0}} \tag{7}$$

and

$$\hat{I}_s^G = \frac{\exp\left[\frac{1}{N_s}\sum_{i=1}^{N_s}\widehat{\ln(p_{is})}\right]}{\exp\left[\frac{1}{N_0}\sum_{i=1}^{N_0}\widehat{\ln(p_{i0})}\right]}, \tag{8}$$

which are consistent estimators of I_s^A and I_s^G provided that the predictors $\hat{p}_{it}$ and $\widehat{\ln(p_{it})}$ are consistent estimators for $E(p_{it})$ and $E[ln(p_{it})]$, respectively. As shown later in the paper, under suitable assumptions, the *hedonic* estimators $\hat{I}_s^A$ and $\hat{I}_s^G$ may be straightforwardly decomposed into quality and QAPIs.

2.2 The hedonic function

Most of the specifications that have been used for the hedonic function in empirical studies differ essentially in the form under which the explanatory variables enter the model (e.g., logs, squares, and interaction terms), while the dependent variable appears either in levels or in logarithms, and the RHS is typically linear in the parameters. In this paper, we do not discuss the first issue, because, for our purposes, the exact specification of the explanatory variables is irrelevant in the sense that any function of the asset characteristics is easily accommodated by the procedures proposed in the next sections to compute AHPI and GHPI.[3] In contrast, as shown later in the paper, both the

[3]Because of this, next, for simplicity, we use broadly the term ‘log-linear’ to denote any regression model that considers logged prices as the dependent variable and uses a specification linear in the parameters for the RHS (e.g., log–log, semi-log, and translog models).

form of the dependent variable and the specification of the RHS of the hedonic function affect decisively the construction of hedonic price indexes.

Given that prices are strictly positive, in this paper, we focus on the following two specifications for hedonic functions: the log-linear model

$$\ln p_{it} = x_{it}\beta_t + u_{it}; \tag{9}$$

and the exponential model

$$p_{it} = \exp\left(x_{it}\beta_t^* + u_{it}^*\right), \tag{10}$$

where u_{it} (u_{it}^*) is the error term and β_t (β_t^*) is a $(k+1)\times 1$ vector of parameters, with elements $\beta_{t,j}$ ($\beta_{t,j}^*$). The parameter $\beta_{t,j}$ ($\beta_{t,j}^*$) is often interpreted as the implicit marginal price for (some function of) characteristic $X_{t,j}$ and is allowed to change over time.[4] While the log-linear model 9 has been widely used in the construction of hedonic price indexes, the exponential model 10 has been rarely applied in the hedonic literature.[5] In fact, when the dependent variable of the hedonic function is chosen to be the price itself, it has been much more common to use the linear model $p_{it} = x_{it}\beta_t^{**} + u_{it}^{**}$, which, however, may generate negative price estimates in applied work.[6]

In a non-stochastic form (i.e., without an error term), models 9 and 10 would represent exactly the same relationship between p_{it} and x_{it}. However, because of the presence of the stochastic error terms u_{it} and u_{it}^*, the two models are not equivalent, as the former requires the assumption $E(u_{it}\,|\,x_{it})=0$, while the latter assumes $E\left[\exp\left(u_{it}^*\right)\middle|x_{it}\right] = 1$. As is well known, neither of those assumptions imply the other, i.e., $E(u_{it}\,|\,x_{it})=0 \Rightarrow E[exp(u_{it})\,|\,x_{it}]\neq 1$ and $E\left[\exp\left(u_{it}^*\right)\middle|x_{it}\right] = 1 \Rightarrow E\left(u_{it}^*\middle|x_{it}\right)\neq 0$.

3 Links between the price index formula and the form of the dependent variable of the hedonic function

This section examines in detail how AHPI and GHPI may be consistently estimated when the dependent variable of the hedonic function is either the price or the logged price.

3.1 *Links in the framework of arithmetic indexes*

The analysis that follows is made first considering a log-linear hedonic function and then for the exponential case. In the end, the conclusions are extended to any hedonic function that uses the logged price or the price as dependent variable.

3.1.1 *The case of a log-linear hedonic function*

Using a log-linear hedonic function as basis for computing an AHPI has been a very popular approach in the hedonic literature; see *inter alia* Coulson (2012), Dorsey, Hu,

[4]See, however, Pakes (2003), who argued that there is no obvious interpretation for the parameters of the hedonic function given that it represents a reduced form model.
[5]See Wooldridge (1992) for one such rare application.
[6]See *inter alia* the application by Hill and Melser (2008) for the housing market, who had to drop dwellings with negative price predictions before computing price indexes.

Mayer and Wang (2010), Hill (2013), Malpezzi, Chun and Green (1998), and Triplett (2006). However, while the estimation of a log-linear hedonic function yields directly consistent estimates for the logarithm of the asset price, $\widehat{\ln(p_{it})} = x_{it}\hat{\beta}_t$ (see Equation 9), not for the price itself, the computation of an AHPI requires consistent estimates of prices, not logged prices. Moreover, because of the stochastic nature of hedonic functions, the antilog of $\widehat{\ln(p_{it})}$, $\exp\left[\widehat{\ln(p_{it})}\right] = \exp\left(x_{it}\hat{\beta}_t\right)$, is not in general a consistent estimator of $E(p_t \mid x_{it})$. Indeed, the log-linear hedonic function 9 implicitly assumes that $p_{it} = \exp(x_{it}\beta_t + u_{it})$, i.e.,

$$E(p_{it}|x_{it}) = \exp(x_{it}\beta_t)E[\exp(u_{it})|x_{it}], \tag{11}$$

where, in general, $E[\exp(u_{it}) \mid x_{it}] \neq 1$; see Section 2.2. Therefore, in the log-linear context, consistent estimates of asset prices require inevitably an estimate of $E[exp(u_{it}) \mid x_{it}]$.

Let $\mu_{it} \equiv E[\exp(u_{it}) \mid x_{it}]$ and assume that

$$\mu_{it} = g\left(x^*_{it}\alpha_t\right), \tag{12}$$

where $g(\cdot)$ may be a non-linear function, x^*_{it} is some function of x_{it}, and α_t is a vector of parameters. Assuming, for the moment, that $g(\cdot)$ is a known function and that a consistent estimator for μ_{it}, $\hat{\mu}_{it} = g\left(x^*_{it}\hat{\alpha}_t\right)$, is available, a consistent estimator of I^A_s of Equation 3 is given by

$$\hat{I}^A_s = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s}\exp\left(x_{is}\hat{\beta}_s\right)g\left(x^*_{is}\hat{\alpha}_s\right)}{\frac{1}{N_0}\sum_{i=1}^{N_0}\exp\left(x_{i0}\hat{\beta}_0\right)g\left(x^*_{i0}\hat{\alpha}_0\right)}. \tag{13}$$

Clearly, the only case where the naive estimator $\exp\left(x_{it}\hat{\beta}_t\right)$ for p_{it} can be used for consistent estimation of I^A_s occurs when μ_{it} is constant across assets and over time ($\mu_{it} = \mu$).

Expression 13 is decomposed into quality and quality-adjusted price components as follows:

$$\hat{I}^A_s = \underbrace{\frac{\frac{1}{N_s}\sum_{i=1}^{N_s}\exp\left(x_{is}\hat{\beta}_b\right)g\left(x^*_{is}\hat{\alpha}_b\right)}{\frac{1}{N_0}\sum_{i=1}^{N_0}\exp\left(x_{i0}\hat{\beta}_b\right)g\left(x^*_{i0}\hat{\alpha}_b\right)}}_{\hat{I}^{A_q}_s}\underbrace{\frac{\frac{1}{N_a}\sum_{i=1}^{N_a}\exp\left(x_{ia}\hat{\beta}_s\right)g\left(x^*_{ia}\hat{\alpha}_s\right)}{\frac{1}{N_a}\sum_{i=1}^{N_a}\exp\left(x_{ia}\hat{\beta}_0\right)g\left(x^*_{ia}\hat{\alpha}_0\right)}}_{\hat{I}^{A_q}_s}, \tag{14}$$

where $\hat{I}^{A_q}_s$ and $\hat{I}^{A_p}_s$ are consistent estimators for, respectively, $I^{A_q}_s$ and $I^{A_p}_s$ of Equation 5. From Equation 14, it is clear that in the scale of interest for the construction of AHPI, the implicit price of each characteristic is a function of both $\hat{\beta}_t$ and $\hat{\alpha}_t$.[7] Thus, the estimation of AHPI based on log-linear hedonic functions requires, in general, the

[7]Hence, if one is interested in testing whether the implicit prices have changed significantly between two periods, the traditional practice of applying a Chow test for assessing the null hypothesis of equal β_t coefficients in the two periods is not enough in this context: the constancy of the parameters α_t must also be tested.

availability of consistent estimates of α_0 and α_s, which, by turn, require the specification of the $g(\cdot)$ function in Equation 12. Instead of making a direct functional form assumption for $g(\cdot)$, it has been much more common to make further assumptions on the distribution of the error term u_{it}, which imply a specific form for $g(\cdot)$.

One possibility consists of assuming that u_{it} is homoskedastic, which implies that $E[\exp(u_{it}) \mid x_{it}]$ does not depend on x_{it}; see Duan (1983). Under this assumption, a consistent estimator of μ_{it} is given by Duan's (1983) smearing estimator, which consists of estimating the unknown error distribution by the empirical distribution function of the OLS residuals of the log-linear model and then taking expectations with respect to that distribution:

$$\hat{\mu}_{it} = \frac{1}{N_t} \sum_{i=1}^{N_t} \exp(\hat{u}_{it}). \tag{15}$$

Alternatively, it may be assumed that u_{it} has a normal distribution with a variance of a known form, $u_{it} \sim \mathrm{N}\left(0, x_{it}^{*}\alpha_t\right)$. As is well known, this implies that $exp(u_{it})$ has a log-normal distribution, with mean given by

$$\mu_{it} = \exp\left(0.5 x_{it}^{*}\alpha_t\right). \tag{16}$$

In this case, an estimate of α_t can be obtained by regressing the squared OLS residuals of the log-linear model on x_{it}^{*}. From now on, we use the term 'normal-smearing estimator' to denote the estimator computed according to Equation 16.

Many authors in the hedonic price literature are aware of the need for applying bias corrections when computing AHPI from log-linear hedonic functions. Clearly, most authors prefer to apply the normal-smearing estimator (e.g., Coulson, 2012; Dorsey, Hu, Mayer and Wang, 2010; Malpezzi, Chun and Green, 1998; Pakes, 2003; Triplett, 2006) rather than the smearing correction (García and Hernández, 2007, seem to be the only authors to have used this estimator), although all of them assume homoskedasticity. Typically, the assumptions underlying the application of the chosen bias correction are not discussed, much less are they tested. Moreover, some authors still do not apply any bias correction in this context, either because it is considered irrelevant or because practitioners are simply not aware of it.[8] The Monte Carlo study in Section 6 illustrates the important biases that may arise if such bias corrections are not implemented.

3.1.2 The case of an exponential hedonic function

As shown in the previous section, the computation of AHPI based on log-linear hedonic functions requires an estimate of μ_{it}. On the other hand, the estimation of

[8]Actually, many authors seem to confuse the bias corrections analyzed in this section, which are necessary for obtaining *consistent* predictors for asset prices, with those discussed by Goldberger (1968), which aim only at reducing the *finite sample bias* of those predictors and, thus, are not important asymptotically. See *inter alia* the recent excellent survey by Hill (2013) on QAPI for residential housing, which, however, discusses only Goldberger's (1968) finite sample bias correction but cites several papers, including a previous version of this paper, which actually address the other, more important, bias correction.

μ_{it} by simple methods requires some stringent assumptions on the distribution of the error term, which, *a priori*, there is no reason to believe will hold with actual data.[9] In this section, we investigate whether the calculation of AHPI is simpler when the hedonic function has an exponential specification.

Assume that the data generating process (DGP) of asset prices is appropriately described by the exponential hedonic function 10, with $E\left[\exp\left(u^*_{it}\right)\middle|x_{it}\right] = 1$. Assume also that the researcher specifies and estimates that same hedonic function. In this framework, a consistent predictor of asset prices is simply given by $\hat{p}_{it} = \exp\left(x_{it}\hat{\beta}^*_t\right)$. Therefore, a consistent estimator of I^A_s is given by the hedonic estimator $\hat{I}^A_s$ of Equation 7, with $\hat{p}_{it}$ replaced by $\exp\left(x_{it}\hat{\beta}^*_t\right)$, which can be straightforwardly decomposed into quality and QAPIs:

$$\hat{I}^A_s = \frac{\frac{1}{N_s}\sum_{i=1}^{N_s}\exp\left(x_{is}\hat{\beta}^*_s\right)}{\frac{1}{N_0}\sum_{i=1}^{N_0}\exp\left(x_{i0}\hat{\beta}^*_0\right)} = \underbrace{\frac{\frac{1}{N_s}\sum_{i=1}^{N_s}\exp\left(x_{is}\hat{\beta}^*_b\right)}{\frac{1}{N_0}\sum_{i=1}^{N_0}\exp\left(x_{i0}\hat{\beta}^*_b\right)}}_{\hat{I}^{A_q}_s}\underbrace{\frac{\frac{1}{N_a}\sum_{i=1}^{N_a}\exp\left(x_{ia}\hat{\beta}^*_s\right)}{\frac{1}{N_a}\sum_{i=1}^{N_a}\exp\left(x_{ia}\hat{\beta}^*_0\right)}}_{\hat{I}^{A_p}_s}. \tag{17}$$

Clearly, the construction of AHPI based on exponential hedonic functions is much simpler, as there is no need to implement any bias corrections. Nevertheless, because log-linear and exponential regression models are not equivalent, it is important to examine the effects of estimating an exponential regression model in cases where the DGP has a log-linear representation. Consider first the augmented log-linear model that assumes in addition to Equation 9 that

$$\mu_{it} = \exp(x_{it}\alpha_t). \tag{18}$$

Then, from Equation 11, it follows that

$$E\left(p_{it}\middle|x_{it}\right) = \exp[x_{it}(\beta_t + \alpha_t)] = \exp\left(x_{it}\beta^*_t\right), \tag{19}$$

where $\beta^*_t \equiv \beta_t + \alpha_t$. Hence, for our purposes, the addition of assumption 18 to the log-linear model is equivalent to assume from the start that the DGP of asset prices is appropriately described by the exponential hedonic function 10. Therefore, the exponential model 19 produces consistent estimators for $E(p_{it}\,|\,x_{it})$ *even when the true hedonic function has a log-linear form*, provided that assumption 18 holds in the data. It is true that β_t and α_t cannot be identified but that is irrelevant for the computation of AHPI.

As shown next, assumption 18 is by no means heavier than those assumptions made in the previous section to ignore or to simplify the estimation of the μ_{it} in the log-linear context. Let $\alpha_{t,0}$ be the intercept and $\alpha_{t,+}$ be the remaining component of α_t. The bias correction may be ignored only if $\mu_{it} = \mu = \exp(w)$, which, relative to Equation 18, imposes two additional constraints: $\alpha_{t,0} = w$ and $\alpha_{t,+} = 0$. The smearing estimator, by

[9]For example, when working with log-linear hedonic functions in the construction of house price indexes, Fletcher, Gallimore and Mangan (2000), Goodman and Thibodeu (1995), and Stevenson (2004) found asset age-induced heteroskedasticity, which prevents application of the simple smearing estimator.

assuming $\mu_{it} = \mu_t = \exp(\alpha_{t,0})$, is also more restrictive, because it requires that $\alpha_{t,+} = 0$. Relative to the normal-smearing estimator, the augmented log-linear formulation does not require normality of u_{it} but adds the assumption $x_{it}^* = x_{it}$. However, functions of x_{it} can be straightforwardly added to the index function in Equation 18. For example, assume that the true hedonic function is log linear and that Equation 16 reduces to $\mu_{it} = \exp\left(0.5x_{it}^2\alpha_t\right)$. Then

$$E\left(p_{it}\middle|x_{it}\right) = \exp(x_{it}\beta_t)\exp\left(0.5x_{it}^2\alpha_t\right) = \exp\left(z_{it}\delta_t^*\right), \tag{20}$$

where z_{it} is a vector containing the distinct elements of both x_{it} and x_{it}^2, and δ_t^* is a vector of parameters. Therefore, assumption 16 is also easy to accommodate in a standard exponential regression model. Thus, the same assumptions that simplify the calculation of AHPI when the hedonic function is log-linear also ensure that the exponential model yields consistent estimators for asset prices.

3.1.3 The general case

The previous analysis suggests that it is much simpler to compute AHPI based on exponential hedonic functions, which do not require the estimation of any bias corrections, than to use log-linear models, in which case not only is there one additional function to be dealt with (the error variance function) but also it is typically less clear how to specify it. Thus, we may conclude that there is a very convenient link between the computation of AHPI and the specification of exponential hedonic functions. However, using an exponential model is not the only way of ensuring that no bias corrections are required to calculate AHPI. Indeed, it is straightforward to see that simple decompositions of unadjusted price indexes as in Equation 17 are also produced by any other hedonic function that uses the price as dependent variable (e.g., the linear model), *irrespective of the particular specification adopted for the RHS*. It is also evident that with any other form of the dependent variable, it will be necessary to use (variants of) the more complex decomposition 14, *irrespective of the particular transformation adopted for the dependent variable of the hedonic function*. Hence, what is effectively relevant for a simple computation of AHPI is that the dependent variable of the hedonic function is the price itself and not some transformation of it.

Thus, we may conclude the following:

Link 1a: *There exists a link between the computation of AHPI and hedonic functions that consider the untransformed asset price as dependent variable.*

If this link is respected, AHPI may be consistently estimated without applying any bias corrections, and one may focus on the issue of choosing the best specification for the RHS of the hedonic function. To this end, we may use standard functional form tests (e.g., RESET) to assess whether the adopted model is in fact an appropriate specification for $E(p_{it}|x_{it})$ and/or employ non-nested hypothesis tests to assess, e.g., linear and exponential models against each other. We may also, following Pakes (2003), use the adjusted R^2 to decide which hedonic model provides the most exact and accurate price predictions. Therefore, the apparently contradictory positions of many authors discussed in the Section 1 are actually conciliable, because: (i) although by convenience the choice of the dependent variable must be dictated by the choice of the price index, there are no *a priori* restrictions on the specification of the RHS of the hedonic function; and (ii) in all empirical studies where log-linear models have been used to compute AHPI, the (sometimes implicit) assumptions made imply that asymptotically equivalent results are produced by exponential hedonic functions.

3.2 Links in the framework of geometric indexes

Using arguments similar to those put forward in the previous section, it is straightforward to show that the construction of GHPI simplifies considerably when the hedonic function uses logged prices as the dependent variable.[10] For example, assuming that the DGP of asset prices is suitably described by a log-linear model, a consistent estimator of I_s^G of Equation 4 is given by the hedonic estimator of Equation 8, with $\ln(\hat{p}_{it})$ replaced by $x_{it}\hat{\beta}_t$, which can be decomposed as follows:

$$\hat{I}_s^G = \frac{\exp\left(\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_s\right)}{\exp\left(\frac{1}{N_0}\sum_{i=1}^{N_0} x_{i0}\hat{\beta}_0\right)} = \underbrace{\frac{\exp\left(\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_b\right)}{\exp\left(\frac{1}{N_0}\sum_{i=1}^{N_0} x_{i0}\hat{\beta}_b\right)}}_{\hat{I}_s^{G_q}} \underbrace{\frac{\exp\left(\frac{1}{N_a}\sum_{i=1}^{N_a} x_{ia}\hat{\beta}_s\right)}{\exp\left(\frac{1}{N_a}\sum_{i=1}^{N_a} x_{ia}\hat{\beta}_0\right)}}_{\hat{I}_s^{G_p}}. \tag{21}$$

In contrast, any specification for the hedonic function that does not use logged prices as the dependent variable will require the use of bias corrections. Therefore:

Link 1b: *There exists a link between the computation of GHPI and hedonic functions that consider logged asset prices as a dependent variable.*

4 Links between the price index formula, the RHS of the hedonic function, and the estimation method

Another issue that is worth investigating is the relation between the sample ($\bar{I}_s^A$ and $\bar{I}_s^G$) and the hedonic ($\hat{I}_s^A$ and $\hat{I}_s^G$) estimators of unadjusted arithmetic and geometric price indexes that were introduced in Section 2.1. Next, we discuss under which circumstances both types of estimators produce identical estimates of unadjusted price variations.

[10]For details, see the supplementary material available at http://evunix.uevora.pt/~jsr/papers/HedonicLinks-Supplement.pdf.

From Equations 2 and 8, a sufficient condition for ensuring that $\overline{I}_s^G = \hat{I}_s^G$ is given by

$$N_t^{-1}\sum_{i=1}^{N_t}\ln p_{it} = N_t^{-1}\sum_{i=1}^{N_t}\widehat{\ln(p_{it})}. \tag{22}$$

In general, this equality does not hold. However, there is a case in which Equation 22 is satisfied. When the hedonic function has a log-linear specification and the parameters of the model are estimated by OLS, the estimator $\hat{\beta}_t$ for β_t satisfies the following set of orthogonality conditions between the residuals $\hat{u}_{it}$ and the explanatory variables:

$$\sum_{i=1}^{N_t} x_{it}^{'}\hat{u}_{it} = \sum_{i=1}^{N_t} x_{it}^{'}\left(\ln p_{it} - \widehat{\ln p_{it}}\right) = 0. \tag{23}$$

Typically, x_{it} includes an intercept, implying that $\sum_{i=1}^{N_t}\hat{u}_{it} = 0$, and, hence, the averages of both the observed and OLS predicted logged prices are identical, as in Equation 22.

Similarly, equality $\overline{I}_t^A = \hat{I}_t^A$ is only satisfied when

$$\frac{1}{N_t}\sum_{i=1}^{N_t} p_{it} = \frac{1}{N_t}\sum_{i=1}^{N_t}\hat{p}_{it}; \tag{24}$$

see expressions 1 and 7. Assuming a linear hedonic function, then, again, estimation by OLS ensures that condition 24 holds. Assuming the more interesting exponential specification recommended in this paper, there are a variety of alternative methods that may be used for estimating β_t^*, but, as shown next, only one of them ensures that condition 24 is satisfied.

The most common methods for estimating the parameters of an exponential regression model are non-linear least squares (NLS), PPML, and Gamma pseudo-maximum likelihood (GPML). In all cases, the only condition required for consistency is the correct specification of the hedonic function, with the methods differing essentially on the functional form assumed for the conditional variance of $\exp\left(u_{it}^*\right)$ in Equation 10:

$$V\left[\exp\left(u_{it}^*\right)|x_{it}\right] = \tau\exp\left(x_{it}\beta_t^*\right)^{-\rho}, \tag{25}$$

where $\rho = 0$ (GPML), $\rho = 1$ (PPML), or $\rho = 2$ (NLS), and τ is a constant term. As shown by Santos Silva and Tenreyro (2006), the first-order conditions for each estimator are given by

$$\sum_{i=1}^{N_t} x_{it}^{'}\left[p_{it} - \exp\left(x_{it}\hat{\beta}_t^*\right)\right]\exp\left(x_{it}\hat{\beta}_t^*\right)^{\rho-1} = 0, \tag{26}$$

which implies that only when $\rho = 1$ (PPML estimator) are the averages of observed and predicted dwelling prices identical as required by condition 24.[11]

[11]The PPML method is available as a canned command in many econometric packages. For instance, it may be implemented in Stata using one of the following command lines: poisson p_{it} $X_{it,1}\ldots X_{it,k}$, robust or ppml p_{it} $X_{it,1}\ldots X_{it,k}$ (the latter command requires the previous installation of the package ppml.ado; just type 'findit ppml' and follow the links). See Santos Silva and Tenreyro (2011) for details on both implementations of the PPML estimator.

A very useful implication of Equations 22 and 24 is that the process of producing Paasche-type QAPI is substantially simplified. Consider $\hat{I}_s^{A_p}$ of Equation 17, with $a = s$. Denote by $\overline{p}_s$ the arithmetic mean of the actual asset prices in period s. Estimating $\hat{\beta}_t^*$ by PPML, it follows from Equation 24 that $\hat{I}_s^{A_p}$ reduces to

$$\hat{I}_s^{A_p} = \frac{\overline{p}_s}{\frac{1}{N_s}\sum_{i=1}^{N_s} \exp\left(x_{is}\hat{\beta}_0^*\right)}; \tag{27}$$

a similar simplification is available if we use a linear hedonic function and estimate it by OLS. In the case of GHPI, $\hat{I}_s^{G_p}$ of Equation 21, with $a = s$ and $\hat{\beta}_t$ estimated by OLS, may be simplified to

$$\hat{I}_s^{G_p} = \frac{\overline{p}_s}{\exp\left[\frac{1}{N_s}\sum_{i=1}^{N_s} x_{is}\hat{\beta}_0\right]}, \tag{28}$$

where $\overline{p}_s$ now denotes a geometric mean. These simplified Paasche-type price indexes are very attractive for statistical agencies, because they may be computed in a more timely and simple manner: the hedonic function needs to be estimated only at the base period.

Thus, we conclude that:

Link 2: *There exists a link between a linear (exponential) specification for the RHS of the hedonic function and the OLS (PPML) estimation method.*

At this point, it is worth acknowledging that Reis and Santos Silva (2006) have also analyzed some of the issues discussed so far. However, in addition to working in another context (they dealt with weighted indexes for frequently traded assets), there are three crucial differences between our approach and theirs. The first has to do with the definition of a QAPI. Right from the start, Reis and Santos Silva (2006) defined (Paasche) QAPI expressions using expressions similar to Equations 27 and 28, i.e., ratios of observed and predicted prices. Then, they argue that only when (the equivalent of) Links 1 and 2 are simultaneously respected is a basic property of any estimator of QAPI satisfied: the index should equal 1 when $\hat{\beta}_0$ equals $\hat{\beta}_s$. In contrast, because our estimators of QAPI are defined as ratios of predicted prices, they always satisfy that basic property, *even when none of the identified links is respected*. As a consequence, while we talk about *convenient* links, their links are presented as *compulsory*. The second relevant difference is that, because they deal with (the equivalent of) Links 1 and 2 simultaneously, they do not evaluate separately the consequences of not respecting only one of the links. In contrast, we have shown that those consequences are markedly different: while failing to respect Link 1 may lead to inconsistent estimation of QAPI if incorrect assumptions about the error term are made, failing to respect Link 2 only means that the process of producing Paasche QAPI will be more time consuming and that in small samples there may be some deviations between hedonic and sample estimates of unadjusted price variation, which tend to zero asymptotically. Finally, Reis and Santos Silva (2006) focused on the computation of geometric price indexes and, hence, have not explicitly dealt with exponential functions, while one of the main aims of this paper is precisely to show the usefulness of such functions in the hedonic framework.

5 Links in the context of the time dummy variable method

The technique used in the previous sections to obtain the price decompositions given in Equations 17 and 21 is known as the imputation price method, which is the most common and flexible hedonic method because it allows the model parameters to vary freely over time.[12] In contrast to this method, the time dummy variable method assumes that the implicit prices of the asset characteristics are constant across a certain number of time periods. Hence, only one hedonic function needs to be estimated for the whole period, using a sample that comprises observations from all periods.

Let T denote that number of periods, let T_i be a vector of $T-1$ dummy variables whose elements T_{it} $(t=1,\ldots,T-1)$ take the value unity if asset i was sold at period t (and zero otherwise), and let λ (λ^*) be the associated vector of coefficients with elements λ_t (λ_t^*). Let also r_{it} be a vector containing all asset characteristics other than the period of sale and θ (θ^*) be the associated vector of parameters that is assumed to be constant over time. Thus, in the log-linear case, the hedonic function may be written as

$$\ln(p_{it}) = r_{it}\theta + T_{it}\lambda_t + u_{it}, \tag{29}$$

while in the exponential case, it is given by

$$p_{it} = \exp\left(r_{it}\theta^* + T_{it}\lambda_t^* + u_{it}^*\right). \tag{30}$$

Under suitable assumptions, consistent predictors for logged asset prices in periods 0 and s are given by, respectively, $\widehat{\ln(p_{i0})} = r_{i0}\hat{\theta}$ and $\widehat{\ln(p_{is})} = r_{is}\hat{\theta} + \hat{\lambda}_s$, and consistent predictors for asset prices are given by, respectively, $\hat{p}_{i0} = \exp\left(r_{i0}\hat{\theta}^*\right)$ and $\hat{p}_{is} = \exp\left(r_{is}\hat{\theta}^* + \hat{\lambda}_s^*\right)$.

From Equation 21, it follows that the GHPI based on the log-linear function 29 simplifies to

$$\hat{I}_s^{G_p} = \frac{\exp\left(\frac{1}{N_a}\sum_{i=1}^{N_a} r_{ia}\hat{\theta} + \hat{\lambda}_s\right)}{\exp\left(\frac{1}{N_a}\sum_{i=1}^{N_a} r_{ia}\hat{\theta}\right)} = \exp\left(\hat{\lambda}_s\right), \tag{31}$$

which is a well-known result in the hedonic literature and, in fact, the main attractiveness of using the time dummy variable method. For this reason, and because most authors

[12] Note that this is true *even when Link 2 is respected*. Indeed, under Link 2, the hedonic regressions for some periods do not need to be explicitly estimated, but the underlying parameters are still allowed to change freely.

seem to think that a similar result is not possible in the AHPI framework, there is an apparent consensus in the hedonic literature that there is a link between the time dummy variable method, the log-linear hedonic function, and the GHPI, in the sense that only with this specific combination of methods, functions, and indexes is the calculation of QAPI substantially simplified. See, for example, Diewert, Heravi and Silver (2009), Haan (2010), Hill (2013), and Silver and Heravi (2007), which, in their sections dedicated to the time dummy variable method, restrict their attention to GHPI calculated from hedonic functions based on logged prices, and Diewert (2011) and Triplett (2006), which consider a linear hedonic function and conclude that no expression similar to Equation 31 is available in the AHPI framework. However, as shown next, a similar simplification applies to AHPI when used in association with the exponential hedonic function 30. Indeed, from Equation 17, it follows that

$$\hat{I}_s^{A_p} = \frac{\frac{1}{N_a}\sum_{i=1}^{N_a} \exp\left(r_{ia}\hat{\theta}^* + \hat{\lambda}_s^*\right)}{\frac{1}{N_a}\sum_{i=1}^{N_a} \exp\left(r_{ia}\hat{\theta}^*\right)} = \exp\left(\hat{\lambda}_s^*\right). \tag{32}$$

Hence, the link established in the hedonic literature is reformulated as follows:

Link 3: *In the framework of the time dummy variable method, there exists a link between the AHPI (GHPI) and hedonic functions that consider the price (logged price) as dependent variable and use an exponential (linear) specification for the RHS.*

6 Monte Carlo simulation study

We use Monte Carlo methods to compare estimators of (fixed base Paasche-type) AHPI based on choices that do and do not respect Link 1 and/or Link 2.[13] Moreover, we also compare the ability of each estimator to predict asset prices. In order to obtain a realistic scenario for our experiments, the housing dataset of Anglin and Gençay (1996) is used as basis for simulating several patterns of evolution for dwelling prices and characteristics. All experiments are based on 5000 replications.

6.1 Anglin and Gençay's (1996) dataset

All Monte Carlo experiments that follow are based on real data for the Canadian housing market, namely for the city of Windsor. This dataset, which was first analyzed by Anglin and Gençay (1996), consists of 546 observations for the year of 1987 and includes 11 explanatory variables: one continuous regressor, four count

[13]Link 3 is not considered, because failure in respecting it does not cause any further bias relative to the one already implied by failure in respecting Link 1. Indeed, if we take into account Link 1 but not Link 3, that is, if we use logged prices (prices) but not a linear (exponential) specification for the right-hand side of the hedonic function, and then compute geometric (arithmetic) indexes, the only negative consequence of such an approach is the necessity of using the full formula for computing the index instead of the simplified formulas given in Equations 31 and 32.

variables, and six binary regressors. To simplify our investigation, without loss of generality, next we consider only one of each type of explanatory variable, namely the natural logarithm of the lot size of the property in square feet (LOT), the number of bedrooms (BDMS), and a dummy variable, which equals one if the dwelling is located in a preferred neighborhood of the city (REG). In all regressions, the dependent variable is the sale price in Canadian dollars, divided by 100,000, or its logarithm.

Regressing the logarithm of the price on a constant term and the three mentioned explanatory variables produces the following results:

$$\widehat{\ln(p_i)} = -4.809 + 0.460\text{LOT}_i + 0.141\text{BDMS}_i + 0.184\text{REG}_i, \tag{33}$$

with $\hat{\sigma}^2 = 0.075$, where σ^2 is the error term variance under the assumption of homoskedasticity. In order to establish a possible heteroskedasticity pattern, we regressed also the squared residuals, $\hat{u}_i^2$, of Equation 33 on LOT_i and its square, obtaining

$$\hat{u}_i^2 = -0.007\text{LOT}_i + 0.002\text{LOT}_i^2 + \text{error}. \tag{34}$$

Finally, we also estimated the following exponential hedonic function:

$$\hat{p}_{it} = \exp(-4.770 + 0.458\text{LOT}_{it} + 0.147\text{BDMS}_{it} + 0.168\text{REG}_{it}). \tag{35}$$

6.2 *Prediction of dwelling prices*

Before focusing on the comparison of alternative estimators for AHPI, we consider the effects of different assumptions on the error term over the prediction of dwelling prices by four different methods. Three of them consider a log-linear hedonic function, estimated by OLS: the naive OLS estimator, which first estimates $ln(p_{it})$ and then uses its antilog as predictor of p_{it}; the normal-smearing OLS estimator (OLSn), which applies the bias correction 16 to the OLS estimator and assumes a normal-distributed error term with known variance; and the smearing OLS estimator (OLSs), which applies the bias correction 15 to the OLS estimator and assumes a homoskedastic error term. The fourth estimator considers an exponential hedonic function, which is estimated by PPML and includes the variable LOT_{it}^2 as an additional regressor under heteroskedasticity.

In this first set of experiments, we simulate 5000 random samples of size 546 drawn with replacement from the actual sample of regressors. Then, we generate dwelling prices for a single time period using the hedonic function $ln(p_i) = -4.809 + 0.460\,\text{LOT}_i + 0.141\text{BDMS}_i + 0.184\text{REG}_i + u_i$, where u_i was generated from three alternative distributions with mean zero and variance σ_i^2: a normal distribution $N\left(0, \sigma_i^2\right)$, a displaced Gamma distribution $\text{Gamma}\left(\gamma^2\sigma_i^2, \gamma\right) - \gamma\sigma_i^2$, where $\gamma = 1.5$, and a Gumbel distribution $\text{Gumbel}(-0.577216\eta_i, \eta_i)$, where $\eta_i = \sqrt{6\sigma_i^2/\pi^2}$. Regarding the error term variance, we considered both the cases of homoskedasticity ($\sigma_i^2 = \sigma^2$, where σ^2 is either 0.075 or 0.375) and heteroskedasticity ($\sigma_i^2 = -0.007\text{LOT}_i + c\text{LOT}_i^2$, where $c = 0.002$ or 0.01).

Following Mood, Graybill and Boes (1974, pp. 540–543), for the normal, Gamma, and Gumbel cases, $E[\exp(u_{it})]$ is given by, respectively, $\exp\left(0.5\sigma_{it}^2\right)$, $[\gamma/(\gamma-1)]^{\gamma^2\sigma_{it}^2}\exp\left(-\gamma\sigma_{it}^2\right)$, and $\Gamma(1-\eta_{it})\exp(-0.577216\eta_{it})$. Therefore, in the normal and Gamma cases,

$$E(p_i|x_i) = \exp\left(x_i\beta + \varphi\sigma_i^2\right) = \exp(z_i\delta^*),$$

where $\varphi = 0.5$ (normal) or $\varphi = \gamma^2 ln[\gamma/(\gamma-1)] - \gamma$ (Gamma) and $z_i = x_i$ (homoskedasticity) or $z_i = \left(x_i, \text{LOT}_i^2\right)$ (heteroskedasticity). In contrast, for the Gumbel case,

$$\begin{aligned} E(p_i|x_i) &= \exp\{x_i\beta - 0.577216\eta_i + \ln[\Gamma(1-\eta_i)]\} \\ &= \exp\left\{x_i\beta - \frac{\sqrt{6}}{\pi}\sigma_i + \ln\left[\Gamma\left(1 - \frac{\sqrt{6}}{\pi}\sigma_i\right)\right]\right\}, \end{aligned}$$

which simplifies to $exp(z_i\delta^*)$ only in case of homoskedasticity and has a non-standard representation otherwise.

Figure 1 displays the average predictions across replications yielded by each estimator for dwellings, which have three bedrooms and are not located in a preferred neighborhood of the city and whose lot size ranges, in steps of 50, from 1650 to 15,600 ft^2. The first two values represent the mode for the variables BDMS and REG, while the boundary values chosen for the lot size are its minimum and maximum values in the sub-sample of dwellings for which (BDMS,REG) = (3,0). Figure 1 displays also the expected value of dwelling prices (denoted as True), i.e., the values obtained from Equation 11 using the true vector β_0, the original data for the regressors, and the correct form for $E[\exp(u_{it}) \mid x_{it}]$.

As shown by Figure 1, none of the OLS estimators are able to yield consistent predictions of dwelling prices across the whole set of experiments. The naive OLS

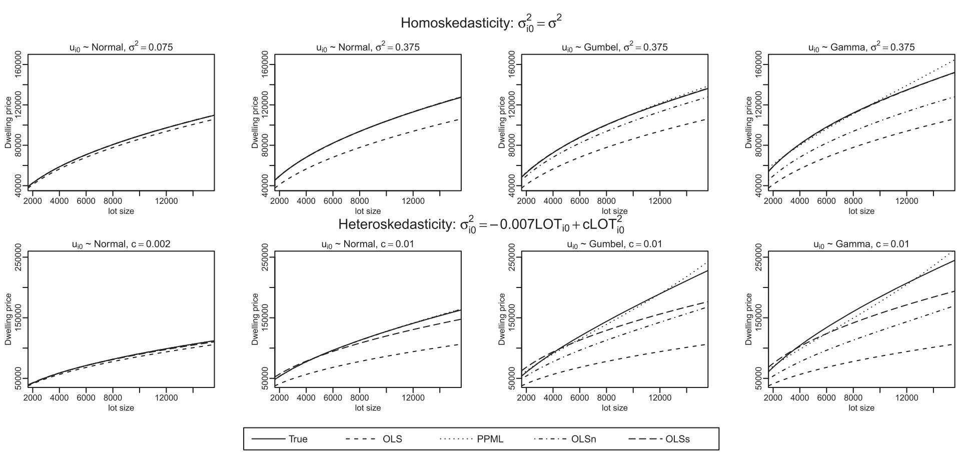

Fig. 1. Alternative methods for predicting dwelling prices ($N_0 = 546$).

estimator produces biased predictions in all cases, with the bias increasing as $E[exp(u_i)|x_i]$ increases from 1.038 (normal, $\sigma_i^2 = 0.075$) to 1.206 (normal, $\sigma_i^2 = 0.375$), 1.289 (Gumbel), and 1.440 (Gamma). The OLSn estimator yields inconsistent predictions whenever u_i has a non-normal distribution. Finally, the smearing estimator fails in consistently predicting dwelling prices under heteroskedasticity, overpredicting the price for low values of the lot size and underpredicting it for high values.

In contrast, the PPML estimator performs well in all experiments, even in the case of a Gumbel-distributed error term, where the exponential regression model estimated is not well specified. Note that the small bias displayed sometimes by the PPML estimator for larger values of LOT is a small sample issue, disappearing asymptotically, as can be confirmed in Figure 2, which is based on 5000 samples of 5460 dwellings drawn with replacement from the original sample. In contrast, the biases of the other estimators do not vanish as the sample size becomes larger.

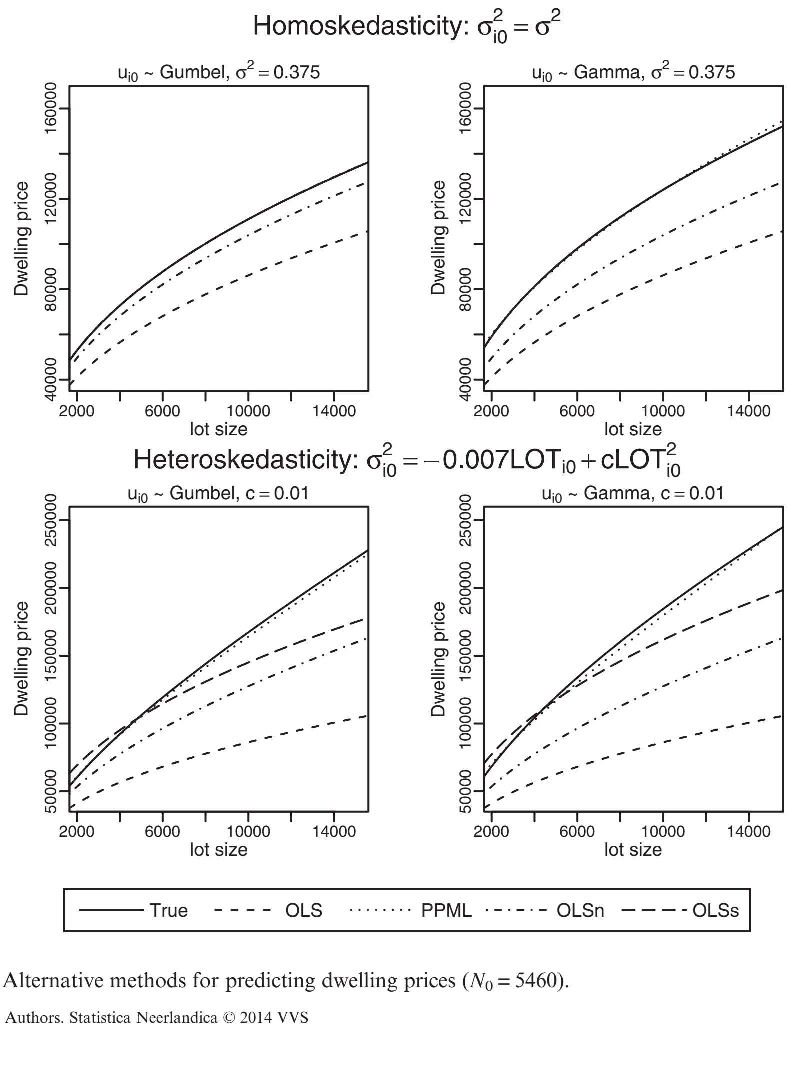

Fig. 2. Alternative methods for predicting dwelling prices ($N_0 = 5460$).

6.3 Link 1

6.3.1 Experimental design

Now, we evaluate the performance of the same four methods when the aim is the estimation of AHPI. The following model was used to generate the dwelling prices in each of the $t = 0, \ldots, 20$ periods that this study comprises:

$$\ln(p_{it}) = \beta_{t,0} + \beta_{t,1}\text{LOT}_{it} + \beta_{t,2}\text{BDMS}_{it} + \beta_{t,3}\text{REG}_{it} + u_{it}. \quad (36)$$

Based on Anglin and Gençay's (1996) dataset, we set $\beta_0' = [-4.809, 0.460, 0.141, 0.184]$. For the remaining periods, $\beta_t = \beta_{t-1}(1 + \Delta\beta_t)$, $t \geq 1$, where the four elements of $\Delta\beta_t$ are drawn independently from a normal distribution with mean zero and variance 0.0001 (design A) or 0.0001/50 (design B). To generate u_{it}, we considered the same three distributions of the previous experiment, with the error term variance being now given by σ_{it}^2. Three distinct patterns were considered for σ_{it}^2: (i) $\sigma_{it}^2 = 0.075$ (homoskedasticity); (ii) $\sigma_{it}^2 = \sigma_t^2 \in [0.075, 0.375]$ (time-varying error variance), either because $\sigma_t^2 = 0.075 + 0.015t$ or is randomly drawn from a uniform distribution on that interval; and (iii) $\sigma_{it}^2 = -0.007\text{LOT}_{it} + c_t\text{LOT}_{it}^2$ (heteroskedasticity), where $c_t \in [0.002, 0.010]$, with $c_t = 0.002 + 0.0004t$ or drawn from a uniform distribution on the mentioned interval.

Monte Carlo samples of size N_t of dwelling characteristics for period 0 were randomly drawn, with replacement, from the original dataset of 546 observations, with N_t being drawn from a uniform distribution with limit points 250 and 500 in order to mimic the fact that with actual data the sample size typically differs across periods. For periods $1, \ldots, 20$, the samples were generated in two steps. In the first step, 'base samples' of size 546 were constructed. In the second step, samples of size N_t were randomly drawn, with replacement, from the base samples. To construct the base samples, first, the dwellings in the original sample where sorted according to their actual sale prices. Then, we constructed four strata, where the first stratum contains the 25% cheapest dwellings, the second comprises the next 25%, and so on. Let f_t be a four-element vector of probabilities assigned to each stratum. We next drew f_t from a Dirichlet distribution with parameter $\varsigma_t = \varphi f_t^B$, where $\varphi = 5$ is a precision parameter, $f_t^B = f_{t-1}^B + \Delta f_t^B$ is the expected value of f_t, $\Delta f_t^B = [-0.01, 0, 0.005, 0.005] * t$, and $f_0^B = [0.25, 0.25, 0.25, 0.25]$.[14] Finally, for each period, we generated a base sample, drawing with replacement from the original dataset a stratified sample based on f_t. Experiments involving tenfold samples were also performed, in which case the same procedures were applied to generate the Monte Carlo sample, but only after replicating the original sample ten times.

To illustrate the main practical characteristics of the experimental designs simulated, Figure 3 displays unadjusted and quality-adjusted arithmetic price indexes, as well as the associated quality index, for both designs A and B when the error term

[14]See Murteira and Ramalho (2013) for details on the particular Dirichlet distribution considered.

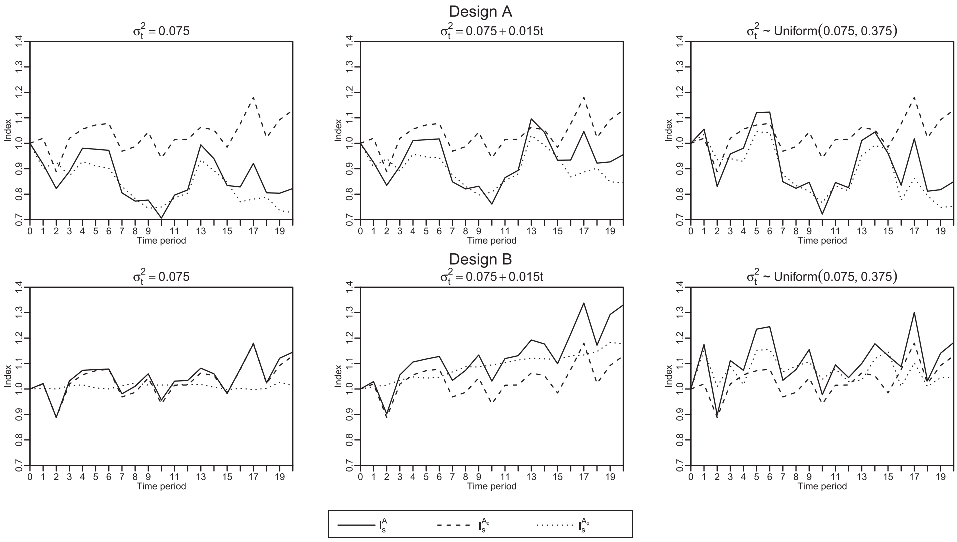

Fig. 3. Population arithmetic price indexes – $u_{it} \sim N(0, \sigma_t^2)$.

has a normal distribution and its variance is defined according to the error variance patterns (i) and (ii) defined earlier. These are fixed base indexes and represent 'population' indexes, as they were calculated using the base samples, the true β_t parameters, and the known bias correction $E[exp(u_{it}) | x_{it}]$.

As Figure 3 shows, the pure price evolution is quite distinct in designs A and B, being much more irregular and displaying much larger absolute variations in the former case. Note also that, although the parameters of the hedonic function are kept fixed across error variance patterns, both the unadjusted and QAPIs vary from experiment to experiment because of different assumptions on the error term variance. For example, consider the three graphs of the second row of Figure 3. In the first graph, $I_s^{A_p}$ changes very little over time. In the second graph, as the result of an increasing variance of the error term, at a constant rate, $I_s^{A_p}$ also increases at a relatively constant rate. Finally, in the third graph, because of the random nature of σ_{it}^2, the time trajectory of $I_s^{A_p}$ is much less regular. This illustrates clearly the need for implementing bias corrections in cases where the hedonic function is specified in a scale that is not the one of interest for calculating the index: when the hedonic function is log linear, while $ln(p_{it})$ and GHPI do not change as a result of a variation of σ_{it}^2, p_{it} and AHPI do change.

6.3.2 Results

Figure 4 reports the main results obtained for alternative estimators of AHPI. The first two rows consider the case of a time-varying error variance, while in the last two rows the error term is heteroskedastic.[15] On the other hand, Table 1 presents

[15]The results for the case where the error term variance is constant over time and across dwellings are omitted because, in such a case, the four estimators yield consistent (and indistinguishable) estimators for $I_s^{A_p}$.

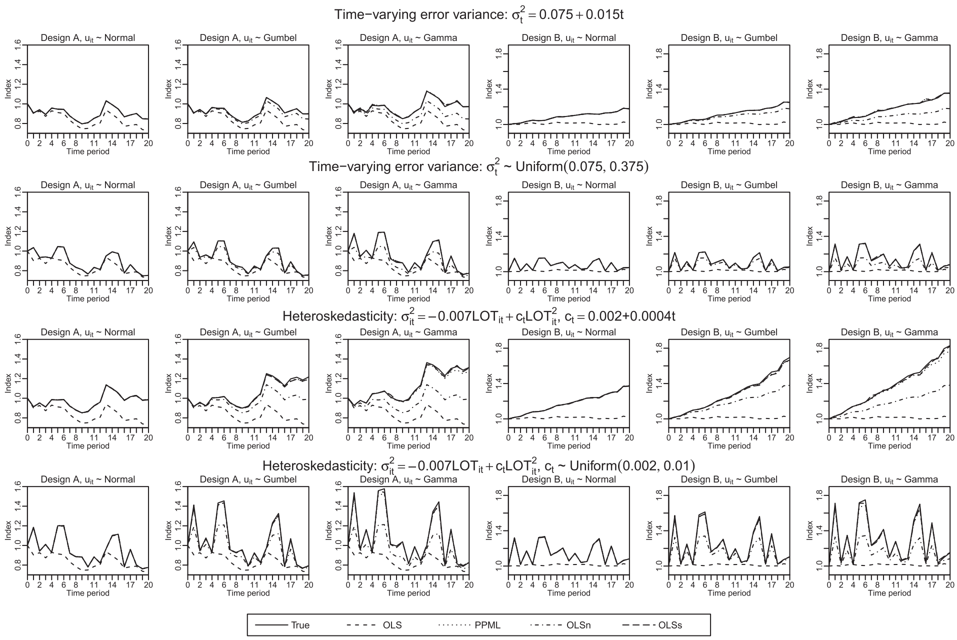

Fig. 4. Alternative estimators of quality-adjusted arithmetic price indexes.

Table 1. Alternative estimators of quality-adjusted arithmetic price indexes: annual growth rates (design A; % bias)

Error distribution	Mean absolute bias				Maximum absolute bias			
	OLS	PPML	OLSn	OLSs	OLS	PPML	OLSn	OLSs
Time-varying error variance: $\sigma_t^2 = 0.075 + 0.015t$								
Normal	0.7	0.0	0.0	0.0	0.8	0.1	0.1	0.1
Gumbel	1.0	0.1	0.3	0.1	1.2	0.2	0.5	0.2
Gamma	1.4	0.6	0.7	0.6	1.6	2.1	1.0	2.1
Time-varying error variance: $\sigma_t^2 \sim \text{Uniform}(0.075, 0.375)$								
Normal	6.5	0.0	0.0	0.0	14.1	0.1	0.1	0.1
Gumbel	9.0	0.0	2.5	0.1	20.5	0.1	5.6	0.2
Gamma	12.6	0.4	6.1	0.5	29.2	1.3	13.1	1.5
Heteroskedasticity: $\sigma_{it}^2 = -0.007\text{LOT}_{it} + c_t\text{LOT}_{it}^2$, $c_t = 0.002 + 0.0004t$								
Normal	1.5	0.1	0.1	0.1	2.2	0.2	0.1	0.3
Gumbel	2.5	0.2	0.2	0.3	4.6	0.8	2.5	1.1
Gamma	2.9	0.3	1.4	1.1	4.2	1.1	2.1	2.8
Heteroskedasticity: $\sigma_{it}^2 = -0.007\text{LOT}_{it} + c_t\text{LOT}_{it}^2$, $c_t \sim \text{Uniform}(0.002,0.01)$								
Normal	12.9	0.1	0.1	0.2	29.1	0.1	0.1	0.7
Gumbel	21.1	0.5	7.7	0.8	53.1	1.7	18.2	1.9
Gamma	25.8	1.5	12.1	1.3	64.9	3.6	27.2	3.6

annual growth rates for the case of design A, reporting the percentage mean and maximum absolute bias of each estimator over the 20 time periods considered (for design B, the results are very similar).

As expected, in these circumstances, any estimates of $I_s^{A_p}$ based on the naive OLS estimator are inconsistent, as Figure 4 clearly shows. Moreover, annual growth rates calculated with basis on $I_s^{A_p}$ may originate mean and maximum absolute biases of 25.8% and 64.9%, respectively (see the last row of Table 1). In fact, the estimates of $I_s^{A_p}$ produced by OLS are independent of the value of σ_{it}^2, while all the other estimators, automatically (PPML) or through a bias correction (OLSn and OLSs), incorporate the effect of a varying error term variance on the untransformed price scale. Note also that while the bias of OLS, in terms of annual growth rates, is not that large when the error term variance increases at a constant rate (see the first and third panels of Table 1), the cumulated effects in the end of the 20 time periods are huge (maximum bias: 44.5% under heteroskedasticity and a Gamma error term).

On the other hand, large deviations from the normality assumption may induce very large biases in the estimation of AHPI by OLSn (maximum bias: 24.5% under heteroskedasticity and a Gamma error term). Moreover, the bias in terms of annual growth rates may also be substantial, with a mean value 12.1% and a maximum of 27.2% in the Gamma, heteroskedastic case.

Regarding the smearing estimator, its performance is very interesting, producing consistent estimates of $I_s^{A_p}$ even in the case of heteroskedasticity.[16] Similarly, the PPML estimator performed well in all experiments, even in the Gumbel case with heteroskedasticity. However, the variability of the smearing estimator is much larger than that of PPML, especially in small samples and with non-normal error terms, as can be seen in Figure 5, which displays the root mean square errors (RMSE) of the estimators under heteroskedasticity for both $N_t \in \{250,500\}$ and $N_t \in \{2500,5000\}$. Interestingly, the OLSn estimator displays the lowest RMSE when the sample size is small. However, this can be hardly seen as a positive feature of this estimator: it just means that, in cases where the error term has a non-normal distribution, OLSn estimates are concentrated far away from the true price indexes. When the sample size increases, the PPML estimator is the best RMSE performer in most cases.

Overall, the results in this section show the importance of respecting Link 1 and, hence, using an exponential hedonic function. When Link 1 is ignored, it is essential to implement the right bias correction, as the bias of the naive OLS estimator may be huge and the relatively popular OLSn only works well when the error term is normally distributed. Somewhat surprisingly, the smearing estimator seems to be an attractive alternative for computing $I_s^{A_p}$, given its apparent robustness to heteroskedasticity. However, this estimator has the undesirable feature of yielding inconsistent estimates of dwelling prices under heteroskedasticity. Moreover, it displays much more variability than the PPML estimator.

[16]We simulated many other experiments with different heteroskedasticity patterns across dwellings, and in all cases, we failed to find an example in which the OLSs-based hedonic indexes would deviate in a sizable manner from the true value of the index.

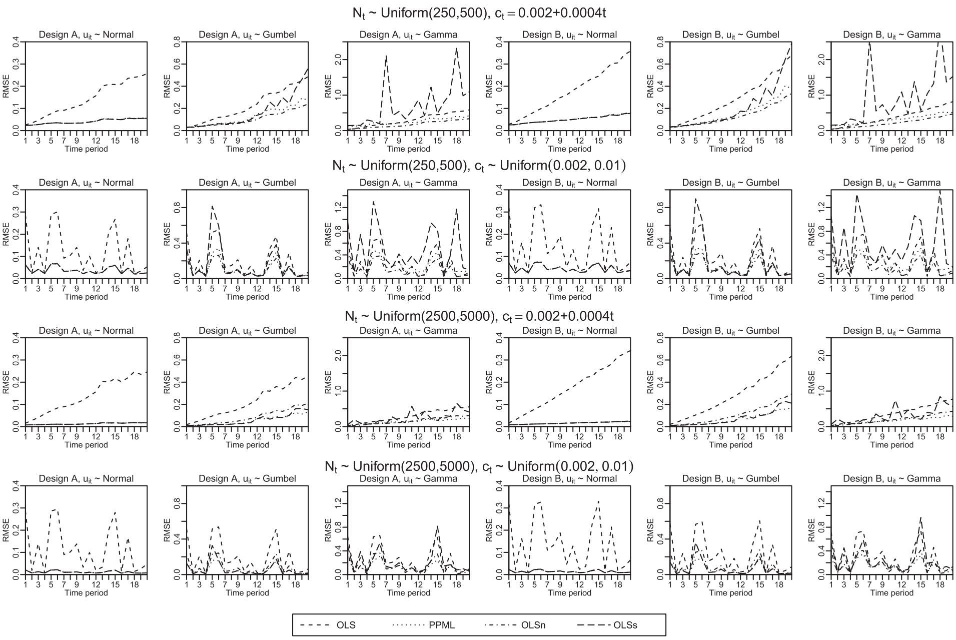

Fig. 5. Root mean square errors – heteroskedastic case: $\sigma^2_{it} = -0.007\text{LOT}_{it} + c_t\text{LOT}^2_{it}$.

6.4 Link 2

In Section 4, we identified a very relevant link between AHPI, exponential hedonic functions, and the PPML method. Next, we investigate whether respecting this link may originate less precise estimates of AHPI in some circumstances by considering two alternative estimation methods to PPML: NLS and GPML. These alternative methods are expected to produce more efficient estimators of the parameters of the hedonic function when the nuisance parameter ρ that appears in the error term variance 25 is close to 0 (GPML) or 2 (NLS).

The following exponential hedonic function is now used to generate dwelling prices:

$$p_{it} = \exp\left(\beta^*_{t,0} + \beta^*_{t,1}\text{LOT}_{it} + \beta^*_{t,2}\text{BDMS}_{it} + \beta^*_{t,3}\text{REG}_{it} + u^*_{it}\right), \tag{37}$$

where, based on Anglin and Gencay's (1996) dataset, we set $\beta^{*'}_0 = [-4.770, 0.458, 0.147, 0.168]$; see Equation 35. We generate $\exp\left(u^*_{it}\right)$ as a log-normal random variable with mean one and variance as in Equation 25, with $\tau = 1$ and $\rho = 1, 0, 1, 2$. The remaining characteristics of these experiments are similar to those of the previous section (design A).

Figure 6 displays 99% confidence intervals and RMSE for alternative estimators of AHPI. Both statistics show clearly that NLS is often much less precise than its competitors, which is a consequence of the extreme values that NLS occasionally yields. These results mimic the erratic behavior of NLS in the estimation of regression coefficients already detected by Manning and Mullahy (2001) and Santos Silva and

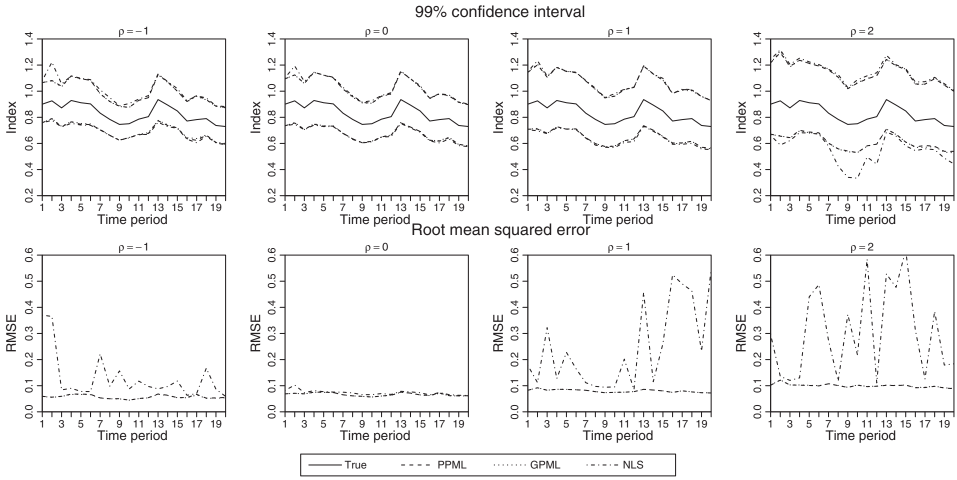

Fig. 6. Alternative quality-adjusted arithmetic price indexes based on exponential hedonic functions.

Tenreyro (2006). Regarding PPML and GPML, no substantial efficiency gains arise from using one or the other estimator, so, given the attractive features of the former estimator discussed before, in general, there will be no reasons for using other estimator than PPML in this context.

7 Conclusion

7.1 Main findings

QAPIs are often computed using hedonic pricing methodologies. In practice, the various choices underlying the estimation of hedonic indexes (price index formula, hedonic function, hedonic method, and estimation method) are usually made in independent ways. In this paper, we have discussed in detail several links between those choices that allow important simplifications in the procedures required to construct a hedonic price index.

The first link identified in the paper concerns the association between AHPI (GHPI) and hedonic functions that use the price (logged price) as dependent variable: only when this link is respected will no bias corrections be necessary to obtain consistent estimates of QAPI. The Monte Carlo study provided clear evidence of the substantial biases that may arise in the construction of AHPI when a log-linear hedonic function is used and wrong assumptions are made on the error term distribution. The other two links discussed in the paper are also very useful, as they allow the computation of QAPI in a more simplified and timely manner. In the context of the imputation price method, the process of producing Paasche-type AHPI is substantially simplified if hedonic functions with linear (exponential) specifications for the RHS are estimated by OLS (PPML). In the framework of the time dummy variable method, the use of an exponential hedonic function allows the computation of AHPI simply as the exponential transformation of a time dummy variable coefficient; so far, such simplification was thought to be valid only for computing GHPI based on log-linear hedonic functions.

Overall, the exponential model, which has rarely been used in the hedonic literature, proves to be more useful to deal with AHPI than the more popular linear model, particularly when estimated by PPML. As the linear model, it avoids the use of bias corrections and allows the simplification of Paasche QAPI if the appropriate estimation method is used. In addition, the exponential model (i) avoids the occurrence of negative predictions for asset prices; (ii) allows the use of the time dummy variable method; and (iii) produces asymptotically equivalent results to those yielded by the popular log-linear hedonic model in the few cases where the bias corrections required to the latter type of model are not too hard to estimate.

7.2 Possible extensions

In this paper, we focused on the construction of AHPI and GHPI. However, there are alternative price index formulas that are commonly used in the computation of QAPI, such as Fisher and Tornqvist indexes.[17] Actually, based on the so-called economic and axiomatic approaches, see Hill (2013), many authors recommend the use of Fisher and Tornqvist indexes over the elementary indexes discussed in this paper. The Fisher QAPI (I_s^F) is given by the geometric mean of Laspeyres and Paasche quality-adjusted AHPI,

$$I_s^F = \sqrt{\frac{E(p_s|x_0)}{E(p_0|x_0)}\frac{E(p_s|x_s)}{E(p_0|x_s)}}, \tag{38}$$

while the Tornqvist QAPI (I_s^T) is given by the geometric mean of Laspeyres and Paasche quality-adjusted GHPI,

$$I_s^T = \sqrt{\frac{\exp\{E[\ln(p_s|x_0)]\}}{\exp\{E[\ln(p_0|x_0)]\}}\frac{\exp\{E[\ln(p_s|x_s)]\}}{\exp\{E[\ln(p_0|x_s)]\}}}. \tag{39}$$

Clearly, given that they are a function of two versions of either I_s^A or I_s^G, the links identified in this paper are also relevant for the computation of Fisher and Tornqvist QAPI.

With a few adaptations, namely the use of weighting schemes, the links identified in this paper also apply to heterogeneous goods frequently transacted. Indeed, based on the work of Reis and Santos Silva (2006), we may conjecture that one just has to use weighted OLS (geometric indexes) or weighted PPML (arithmetic indexes), with the same weights being used in the construction of the price index and in the estimation of the hedonic function.

[17] For a comprehensive text on index number theory, see Balk (2008).

Finally, in many areas of economics, interest lies not in the computation of QAPI but in quality indexes. Clearly, in this case, Link 1 is also relevant. For example, from Equations 14 and 17, it follows immediately that, unless Link 1a is respected, the construction of arithmetic quality indexes requires the use of bias corrections.

Acknowledgements

The authors thank the Editor, the Associate Editor, the referees, and Carlos Brás, Daniel Santos, Erwin Diewert, Marc Francke, João Santos Silva, Rui Evangelista, and Vanda Guerreiro for their valuable suggestions and remarks. Previous drafts have been presented at the 26th Annual Congress of the European Economic Association, Oslo; the 58th World Statistics Congress of the International Statistical Institute (ISI), Dublin; the 18th International Conference on Computing in Economics and Finance, Prague; the 2012 AsRES – AREUEA Joint International Conference, Singapore; and the 7th Annual Meeting of the Portuguese Economic Journal, Covilhã; and seminars at Universidade de Evora, Queen Mary University of London, University of Oporto, University of Otago, and University of Sydney. The authors are pleased to acknowledge financial support from Fundação para a Ciência e a Tecnologia, program FEDER/POCI 2010.

References

Anglin, P. M. and R. Gençay (1996), Semiparametric estimation of an hedonic price function, *Journal of Applied Econometrics* **11**, 633–648.

Ashenfelter, O. and K. Graddy (2003), Auctions and the price of art, *Journal of Economic Literature* **41**, 763–787.

Balk, B. M. (2008), *Price and quantity index numbers: models for measuring aggregate change and difference*, Cambridge University Press, New York.

Campos, N. F. and R. L. Barbosa (2009), Paintings and numbers: an econometric investigation of sales rates, prices, and returns in Latin American art auctions, *Oxford Economic Papers* **61**(1), 28–51.

Chanel, O., A. L. Gérard-Varet and V. Ginsburg (1996), The relevance of hedonic price indices, *Journal of Cultural Economics.* **20**, 1–24.

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

Coulson, N. E. (2012), House price index methodologies, in: S. J. Smith, M. Elsinga, L. Fox-O'Mahony, S. E. Ong and S. Wachter (eds.) *International encyclopedia of housing and home*, Elsevier, Amsterdam, forthcoming.

Diewert, W. E. (2011), Alternative approaches to measuring house price inflation, mimeo.

Diewert, W. E., S. Heravi and M. Silver (2009), Hedonic imputation versus time dummy indexes, in: W. E. Diewert, J. Greenlees and C. R. Hulten (eds.), *Price index concepts and measurement, NBER, Studies in income and wealth*, Vol. 70, University of Chicago Press, Chicago, pp. 161–202.

Dorsey, R. E., H. Hu, W. J. Mayer and H. C. Wang (2010), Hedonic versus repeat-sales housing price indexes for measuring the recent boom-bust cycle, *Journal of Housing Economics* **19**(2), 75–93.

Duan, N. (1983), Smearing estimate: a nonparametric retransformation method, *Journal of the American Statistical Association*, **78**(383), 605–610.

Eurostat (2013), Handbook on residential property price indices (RPPIs), Eurostat Methodologies & Working papers, Publications Office of the European Union. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/hicp/methodology/hps/rppi_handbook.

FLETCHER, M., P. GALLIMORE and J. MANGAN (2000), Heteroskedasticity in hedonic house price models, *Journal of Property Research* **17**(2), 93–108.

FOGARTY, J. J. and C. JONES (2011), The return to wine: a comparison of the hedonic, hybrid, and repeat sales approaches, *Australian Economic Papers* **50**(4), 147–156.

FORTIN, N., T. LEMIEUX and S. FIRPO (2011), Decomposition methods, in: O. Ashenfelter and D. Card (eds.) *Handbook of labor economics*, vol. 4A, Elsevier, Amsterdam, pp. 1–102.

GARCÍA, J. A. B. and J. E. R. HERNÁNDEZ (2007), Housing and urban location decisions in Spain: an econometric analysis with unobserved heterogeneity, *Urban Studies* **44**(9), 1657–1676.

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

GOLDBERGER, A. S. (1968), The interpretation and estimation of Cobb–Douglas functions, *Econometrica* **35**(3/4), 464–472.

GOODMAN, A. C. and T. G. THIBODEAU (1995), Age-related heteroskedasticity in hedonic house price equations, *Journal of Housing Research* **6**(1), 25–42.

HAAN, J. (2004), "Estimating quality-adjusted unit value indices: evidence from scanner data," Paper presented at the Seventh EMG Workshop, Sydney, Australia, December 12-14, SSHRC International Conference on Index Number Theory and the Measurement of Prices and Productivity, Vancouver, June 30-July 3, 2004.

HAAN, J. DE. (2010), Hedonic price indexes: a comparison of imputation, time dummy and 're-pricing' methods, *Jahrbucher fur National Okonomie und Statistik* **230**(6), 772–791.

HILL, R. (2013), Hedonic price indexes for residential housing: a survey, evaluation and taxonomy, *Journal of Economic Surveys* **27**(5), 879–914.

HILL, R. J. and D. MELSER (2008), Hedonic imputation and the price index problem: an application to housing, *Economic Inquiry* **46**(4), 593–609.

MALPEZZI, S., G. H. CHUN and R. K. GREEN (1998), New place-to-place housing price indexes for U.S. metropolitan areas, and their determinants, *Real Estate Economics* **26**(2), 235–274.

MANNING, W. G. and J. MULLAHY (2001), Estimating log models: to transform or not transform?, *Journal of Health Economics* **20**, 461–494.

MOOD, A. M., F. A. GRAYBILL and D. C. BOES (1974), *Introduction to the theory of statistics*, 3rd Ed., McGraw-Hill, Singapore.

MURTEIRA, J. and J. J. S. RAMALHO (2013), Regression analysis of multivariate fractional data, *Econometric Reviews*, forthcoming (DOI: 10.1080/07474938.2013.806849).

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

REIS, H. and J. M. C. SANTOS SILVA (2006), Hedonic indexes for new passenger cars in Portugal (1997–2001), *Economic Modelling* **23**(6), 890–908.

RENNEBOOG, L. D. R. and C. SPAENJERS (2012), Hard assets: the returns on rare diamonds and gems, *Finance Research Letters* **9**(4), 220–230.

SANTOS SILVA, J. M. C. and S. TENREYRO (2006), The log of gravity, *Review of Economics and Statistics* **88**(4), 641–658.

SANTOS SILVA, J. M. C. and S. TENREYRO (2011), Poisson: some convergence issues, *STATA Journal* **11**(2), 207–212.

SILVER, M. and S. HERAVI (2007), The difference between hedonic imputation indexes and time dummy hedonic indexes, *Journal of Business & Economic Statistics* **25**(2), 239–246.

STEVENSON, S. (2004), New empirical evidence on heteroscedasticity in hedonic housing models, *Journal of Housing Economics* **13**(2), 136–153.

TRIPLETT, J. (2006), *Handbook on hedonic indexes and quality adjustments in price indexes: special application to information technology products*, OECD, Paris.

WOOLDRIDGE, J. M. (1992), Some alternatives to the Box–Cox regression model, *International Economic Review* **33**(4), 935–955.

Received: 3 June 2013. Revised: 5 November 2013.