UNI FR
UNIVERSITÉ DE FRIBOURG FACULTÉ DES SCIENCES ÉCONOMIQUES ET SOCIALES
UNIVERSITÄT FREIBURG WIRTSCHAFTS- UND SOZIALWISSENSCHAFTLICHE FAKULTÄT

FOUR ESSAYS ON STATISTICAL PROBLEMS OF HEDONIC METHODS

An Application to Single-Family Houses

THESIS

presented to the Faculty of Economics and Social Sciences at the University of Fribourg (Switzerland)

by

OLIVIER SCHÖNI
from Sumiswald (BE)

in fulfilment of the requirements for the degree of Doctor of Economics and Social Sciences

Accepted by

the Faculty of Economics and Social Sciences
on 18.11.2013 at the proposal of

First Advisor: Prof. Laurent Donzé,
University of Fribourg

Second Advisor: Prof. Eva Cantoni,
University of Geneva

Fribourg, Switzerland 2014

The Faculty of Economics and Social Sciences of the University of Fribourg neither approves nor disapproves the opinions expressed in a doctoral thesis. They are to be considered those of the author.
(Decision of the Faculty Council of 23 January 1990)

This thesis is dedicated to my wife and parents.

ABSTRACT

Over the last ten years, the hedonic approach has been acknowledged as the most appropriate method for addressing the valuation of goods that have a non-constant quality. This thesis is structured in four independent papers that investigate statistical problems related to this approach. The aim is to improve actual knowledge in the hedonic field through an empirical or theoretical approach, and provide results that are useful for researchers and practitioners.
The price index problem in the hedonic context is discussed in the first two papers. The first paper introduces a theoretical framework for estimating hedonic price indices and their confidence intervals. The second paper analyzes the asymptotic properties of the most common hedonic price indices. The third paper focuses on variable selection in hedonic models where multicollinearity is present. In particular, it proposes a new variable selection algorithm that outperforms ordinary automated selection techniques. The fourth paper implements a methodology for comparing the prediction accuracy of two hedonic models.

A statistical analysis, properly conducted,
is a delicate dissection of uncertainties,
a surgery of suppositions.

— M.J. Moroney

ACKNOWLEDGMENTS

First and foremost, I would like to sincerely thank my first advisor, Prof. Laurent Donzé, for his exemplary guidance and advice. Without his help this thesis would not have been possible.
In addition, I would like to extend heartfelt thanks to Prof. Eva Cantoni, who kindly agreed to be my second advisor.
I would like to acknowledge the late Prof. Hans Wolfgang Brachinger, who was initially designated as my first advisor. Before his untimely death, he provided me with the subject of this thesis and the corresponding domain of application.
I would also like to express my sincere gratitude to Michael Beer, who gave me the opportunity to work with him.
Finally, I would like to thank Wüest&Partner for providing the data for the statistical analysis.

Special thanks: Dragana Djurdjevic, Jean-François Emmenegger, Helga Kahr, Christoph Leuenberger, Prof. Jacques Pasquier, Vincent Pochon, Lukas Seger, Daniel Suter.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

OLS Ordinary Least Squares

MBA Mean Balanced Accuracy

AIC Akaike Information Criterion

BIC Bayesian Information Criterion

EPI Elementary Price Indices

HEPI Hedonic Elementary Price Indices

CPI Consumer Price Indices

HC Heteroskedasticity-Consistent

VIF Variance Inflation Factor

RESET Ramsey Regression Equation Specification Error Test

Part I

INTRODUCTION

1

HOUSING GOODS AND THE HEDONIC APPROACH

1.1 THE HEDONIC APPROACH

In the last ten years, the hedonic approach has been acknowledged as the most appropriate method for addressing the valuation of goods that have a non-constant quality. The starting point of this approach is the hedonic hypothesis: Each good is considered a bundle of characteristics, and its price depends solely on these characteristics. The economic theory supporting the hedonic hypothesis dates back to Lancaster (1966) and was successively developed by Rosen (1974).
Let p_i^t and $\mathbf{x}_i^t := (x_{i,1}^t, ..., x_{i,K}^t)$, $i = 1, ..., n$ denote the price of a good i at time t and its corresponding K characteristics. The hedonic hypothesis assumes that the price of a good depends on its characteristics:

$$p_i^t = f^t(x_{i,1}^t, ..., x_{i,K}^t),\ i = 1, ..., n_t,\ t = 1, ..., T,$$

where the function f_t describes how the characteristics interact to build the price. The number of goods observed in period t is denoted by n_t, and it is assumed that $n_t \geqslant K$. Clearly, this analytical relation does not hold in real-world situations, where too many factors participate in the price-building process. To solve this problem, it is assumed that the observed price and characteristics are realizations of random variables related through the following stochastic model

$$p_i^t = f^t(x_{i,1}^t, ..., x_{i,K}^t) + u_i^t,\ i = 1, ..., n_t,\ t = 1, ..., T, \qquad (1)$$

where u_i^t represents a stochastic error term encompassing all the factors not accounted for by the characteristics. The function f^t is usually called the hedonic regression function, or simply the hedonic function. Unfortunately, as pointed out by Rosen (1974), no economic model allows to define an a priori functional form, therefore implying that estimating the hedonic function should be a data-driven process.
Depending on the use of equation (1), hedonic research can be classified in two categories. The first category focuses on

the impact of a specific subset of characteristics on the price. This category has been extensively treated in the hedonic literature, and includes studies attempting to price environmental amenities, measuring the impact of socio-demographic variables, and evaluating policy decisions. This category typically uses an econometric framework, for the purpose is to conduct a ceteris paribus analysis of the variables influencing the price. The classical approach consists of linearly regressing the price of a house on its characteristics, and considers the estimated coefficients implicit prices of the characteristics. Testing procedures are then used to assess the impact of each characteristic on the price of the good.
The second category includes studies aiming to investigate the predictive accuracy of the hedonic model (1). To this end, the traditional ceteris paribus point of view is not required, and a statistical approach using a wide range of estimation techniques is adopted. In particular, the linear hedonic function is often considered too restrictive, and semiparametric and nonparametric estimation techniques are used.
Surprisingly, although prediction is of utmost importance in computing hedonic price indices, this category has played a minor role in the hedonic research field. Few researchers have addressed the problem of predicting prices according to given characteristics, and even fewer have tried to build hedonic price indices for housing goods.
Two major domains in which hedonic methods are applied are high-technology goods and housing goods. In this thesis, we focus on housing goods and, in particular, on single-family houses for the canton of Zurich. In contrast to the international research stream, hedonic research applied to housing goods in Switzerland has been undertaken in the two categories, as described in the next section.

1.1.1 *Hedonic research on housing in Switzerland*

The first research paper considering a hedonic approach using Swiss housing data dates back to Thalmann (1987), who showed differences in the characteristics affecting rents between the city of Lausanne and its district. Grosclaude and Soguel (1992) used a hedonic regression to assess the impact of traffic noise pollution on rents for the city of Neuchatel. A study investigating the effect of traffic noise on rents in residential locations was also conducted by Iten and Maibach (1992) for

the Zurich Agglomeration. In his dissertation Riedi (1992) developed a price index based on a hedonic approach for the canton of Ticino. Studies focusing on hedonic price indices for the real estate market were also carried out by Bender et al. (1994). Bignasca et al. (1996) used a hedonic approach to analyze the price of houses financed by the cantonal bank of Zurich. Hoesli et al. (1997a) and Hoesli et al. (1997b) subsequently compared hedonic price indices with price indices resulting from a repeat sales approach. In a PhD thesis, Scognamiglio (2000) empirically compared traditional real estate assessment methods with the hedonic assessment method, and provided practical instructions for implementing these methods. Hedonic price indices based on rents were constructed by Fahrländer (Fahrländer (2001a) and Fahrländer (2001b)) according to different geographic regions of the Swiss housing market. Salvi et al. (2004) published a technical report for the cantonal bank of Zurich, in which the impact of several characteristics on the house price was analyzed. In their report, a hedonic price index was also computed and then compared to a price index that did not account for quality changes. In 2005, Baranzini and Ramirez (2005) estimated the impact of noise on rents in Geneva. Fahrländer (2006) was the first who used a semiparametric hedonic approach to analyze the prices of condominiums and single-family houses at the nationwide level. In 2008, adding land-specific variables to the classic hedonic model, Baranzini et al. (2008b) assessed how land affects gross monthly rents in the urban areas of Geneva and Zurich.
In 2008, the Swiss Journal of Economics and Statistics dedicated an entire issue to hedonic methods applied to housing in Switzerland. After the introductory paper by Baranzini et al. (2008a), Bourassa et al. (2008) compared the Swiss single-family and condominiums price indices published by the Swiss National Bank with hedonic-based price indices published by IAZI and W&P.[1] Salvi (2008) estimated the impact of airport noise on property prices, taking into account spatial correlation in the analysis. In a research paper, Fahrländer (2008) compared two methods for computing nationwide hedonic price indices for single-family houses and condominiums. In a paper, Banfi et al. (2008) again considered the problem of valuing the effect of environmental disturbances on housing goods. After having

1 Informations-und Ausbildungs-Zentrum für Immobilien (IAZI) and Wüest & Partner (W&P) are two leading real estate consultancy firms in Switzerland.

tested for market segmentation in the rental housing market for Swiss Alpine resorts, Soguel et al. (2008) analyzed whether environmental characteristics were priced differently according to market segmentation. Djurdjevic et al. (2008) used a multilevel model and a classic hedonic model for rents in Switzerland to model housing submarkets and assess the predictive capability of these models. Finally, using a data set on monthly rents for Geneva and Zurich, Baranzini et al. (2008c) established that foreigners were penalized, and paid more than autochthones for a given dwelling quality.
Baranzini and Schaerer (2011) included in a hedonic model characteristics based on a geographic information system to estimate the impact of environmental variables on the Geneva rental market. Bourassa et al. (2011) used the hedonic approach to compute price indices for land value, land leverage, and house value. They estimated an error correction model of house prices and land leverage to determine which variables affect house prices and cause changes in land leverage.

1.2 HEDONIC PRICE INDICES FOR HOUSING GOODS

When hedonic methods and, in particular, hedonic price indices are computed for housing goods, considerable attention must be devoted to the adopted methodology. This because of the magnitude of the investments realized in the housing market. As carefully explained in a recent report published by the OECD (2013), buying a house typically represents the single largest investment in the life of a household and, therefore, a sizeable amount of risk for mortgage lenders. Moreover, researchers and practitioners consider house price indices indicators reflecting the financial stability and the economic activity of a nation. Households may therefore want to use house price indices as wealth indicators and as instruments to make investment decisions, whereas firms could use house price indices to gauge risk exposure and assess the households' borrowing capacity. Additionally, governments could use house price indices as a macroeconomic indicator to help make monetary policy decisions and measure inflation. Recognizing the importance of a sound indicator for the housing market, the Swiss Federal Council has approved the creation of a nationwide price index based on the hedonic approach for the Swiss housing market. The hedonic price index should be introduced in the official statistics in 2017.

To measure price changes in housing goods relative to a base period, price statisticians have defined different sorts of price indices, each possessing specific statistical properties. Although the mathematical definitions of these indices differ greatly, a particular sort of price indices, the Elementary Price Indices (EPI), forms the basis upon which every price index is constructed.

Let $G := \{g_1, ..., g_n\}$ be a set of n goods for which the price behaviour has to be analyzed over the periods $t = 1, ..., T$, and let $\mathbf{p}^t := (p_1^t, ..., p_n^t)'$ be the sample vector containing the prices of the goods belonging to G measured in period t. Simplifying the price index formulae proposed by Beer (2006), two main approaches estimate an elementary price index for a given set of goods G :

$$\widehat{EPI}^{st} = \frac{\hat{\mu}(\mathbf{p}^t)}{\hat{\mu}(\mathbf{p}^s)} \qquad \text{and} \qquad \widehat{EPI}^{st} = \hat{\mu}\left(\frac{\mathbf{p}^t}{\mathbf{p}^s}\right), \tag{2}$$

where $\frac{\mathbf{p}^t}{\mathbf{p}^s} := \left(\frac{p_1^t}{p_1^s}, ..., \frac{p_n^t}{p_n^s}\right)'$ represents the sample vector of the price ratios of the goods in the base period s and in the current period t, respectively. The functional $\hat{\mu}$ is the estimator of the central tendency measure of the observed prices. Interestingly, economic theory provides no guiding model in the choice of the elementary index: The basic index formula and the estimator of the central tendency measure μ must be selected according to the axiomatic and statistical properties of the index. In addition, to be effective, these formulae have to be complemented with a practical description of the economic phenomenon of interest, since the variables chosen for the price index formula depend on the purpose of the price index. Two main choices are effectuated.

First, the objects involved in the phenomena have to be identified, i.e., the set G must be defined according to the research question. For housing goods, an initial distinction is usually made between rental and owned properties. This distinction can subsequently be refined to consider different types of rental and owned properties. For example, owned properties can be divided into residential and commercial properties. In turn, residential properties can be divided into detached houses and condominiums. Another standard partition is made according to geographic regions, where housing markets may behave differently. Very often, a price index is an aggregate of sub-indices, each sub-index corresponding to a different object. These sub-indices usually determine if price changes emanate from a par-

ticular segment and/or good type of the housing market.
Second, a price measure must be associated with the considered objects. Once the considered objects have been defined, specific prices have to be measured. For residential properties, a distinction is usually made between sale and stock prices. The sale price of a house represents the price resulting from the supply-and-demand interaction, whereas stock prices measure the value at a given time. Price indices based on sale prices are mainly used to measure the inflationary pressures that households face, while price indices based on stock prices measure wealth changes and variations in the households' borrowing capacity.
Importantly, the set G and the price measure depend on the purpose of the index, and are independently chosen regarding the mathematical definition of the index. The purpose of the index allows the practical implementation of the mathematical definition (2). Once the purpose of the index has been defined, no multiple uses of the index are thus allowed.
One major methodological problem arises when considering sale prices, and is related to the change in the quality of the good, in our case purchased houses. One implicit assumption in the price index definition (2) is the quality invariance of objects sold during different periods. Whereas we can assume that the quality of the stock of houses remains approximately the same from one month to another due to the small fraction of new houses entering the market, this may not be the case for the basket of purchased houses. Due to extreme house heterogeneity, each house at a given time is considered a unique good, and the observed period-specific prices are greatly influenced by the quality of the purchased houses. The quality problem stems from the intrinsic quality of each single good and the quality of the purchased basket of goods for the entire market. In fact, even if the same basket of houses is purchased between two periods, the quality inherent in each specific good has changed because of time depreciation or refurbishments. The problem of quality change is particularly acute for indices built on the central tendency measures of prices; these indices mix price and quality changes. Even when ratios of median prices are considered, a marked discrepancy is observed with the quality-adjusted price indices (see Bourassa et al. (2008)). To address the quality-variation problem, several hedonic price indices have been proposed in the hedonic literature. The following classification is based on the work of Hill (2012).

1.2.1 *Time-dummy hedonic price indices*

The basic idea of time-dummy hedonic price indices is to not rely on any specific formula from price index theory, but to directly exploit the stochastic version of the hedonic hypothesis (1) to compute quality-adjusted price indices. This intent is achieved by introducing time-dummy variables in equation (2), and considering the coefficients of these dummy variables as quality-adjusted price indices. The standard approach consists of estimating the following model by pooling all cross-sectional data together:

$$\mathbf{p} = f(\mathbf{X}) + \mathbf{D}\gamma + \mathbf{u}, \tag{3}$$

where $\mathbf{p}$ is a price vector with $N := \sum_{t=1}^{T} n_t$ components, and $\mathbf{X}$ is a $N \times K$ matrix containing the characteristics of the housing goods over all time periods. The hedonic function is defined as $f(\mathbf{X}) := (f(\mathbf{x}_1^1), f(\mathbf{x}_2^1), ..., f(\mathbf{x}_N^T))'$. The matrix $\mathbf{D}$ is a $N \times (T-1)$ matrix containing the time-dummy variables: $\mathbf{D}_{it} = 1$ if observation i occurs in period t, and $\mathbf{D}_{it} = 0$ otherwise, with $i = 1, ...N$ and $t = 2, ...T$. The coefficients $\gamma = (\gamma_2, ..., \gamma_T)'$ represent the quality-adjusted price indices. The stochastic error term is represented by the N-vector $\mathbf{u}$.

Although conceptually simple, and avoiding the problem of choosing among different price index formulae, time-dummy indices suffer from severe drawbacks. First, because an explicit price index formula is not used, the axiomatic properties of the computed price indices cannot be used. Second, the hedonic function f is the same across all time periods, implying that the price building-process is constant over time. This may be true when housing goods are considered over several quarters, but the hypothesis seems too restrictive when a time horizon of several years is considered. Third, the computed indices γ may be very sensitive to the functional form f and the estimation method used to estimate equation (3), thus casting doubt on the reliability of such indices. Finally, as observations from new time periods are added to the pooled data set, the previously computed indices may change magnitude, possibly invalidating the results obtained in previous periods. Due to these drawbacks, time-dummy hedonic price indices are not considered in the present thesis.

1.2.2 *Single and double imputed hedonic price indices*

In contrast to time-dummy indices, single and double imputed hedonic price indices address the variation quality problem by modifying the usual price index formulae. Since the price index formulae in (2) assume that prices for the same set of goods G are compared between a base period s and a time period t, single imputed hedonic indices use stochastic equation (1) to predict the price of the goods in G for a given time period according to the price-building process estimated in the other time period. The resulting Hedonic Elementary Price Indices (HEPI) are given by:

$$\widehat{\text{HEPI}}^{st} = \frac{\hat{\mu}(\hat{\mathbf{p}}^t)}{\hat{\mu}(\mathbf{p}^s)} \qquad \text{and} \qquad \widehat{\text{HEPI}}^{st} = \hat{\mu}\left(\frac{\hat{\mathbf{p}}^t}{\mathbf{p}^s}\right). \tag{4}$$

In contrast to single imputed hedonic price indices, where observed prices are used in one of the two periods, double imputed hedonic price indices also estimate the prices in the base period:

$$\widehat{\text{HEPI}}^{st} = \frac{\hat{\mu}(\hat{\mathbf{p}}^t)}{\hat{\mu}(\hat{\mathbf{p}}^s)} \qquad \text{and} \qquad \widehat{\text{HEPI}}^{st} = \hat{\mu}\left(\frac{\hat{\mathbf{p}}^t}{\hat{\mathbf{p}}^s}\right). \tag{5}$$

Single and double imputed hedonic price indices offer important advantages over time-dummy price indices. In each time period, a specific hedonic function f_t is estimated, thus allowing the imputed price index to take into account a possible modification of the price-building process occurring in the housing market. Moreover, since in each period the estimation is carried out independently, adding new time periods does not influence the price index values previously computed. Finally, due to the explicit price index formulae used, the axiomatic approach from price index theory can be used.

These hedonic price indices, however, inherit all the estimation problems related to equation (1), and are subject to the price index problem plaguing price index theory. In the author's opinion, however, imputed hedonic price indices provide a better approach to the quality variation problem and therefore represent the chosen approach in the present dissertation. The problems related to these price indices are reviewed in Section 1.3.

1.2.3 *Characteristic hedonic price indices*

Characteristic hedonic price indices are similar to imputed indices, since in this case the hedonic function is also independently estimated in each time period. Instead of predicting the prices for the whole set G according to the price-generating process estimated in the other time period, however, a single good g^* representative of the set G is chosen. The price behaviour of this characteristic good is then compared across the two time periods:

$$\widehat{\text{HEPI}}^{st} = \frac{\hat{p}^t(g^*)}{\hat{p}^s(g^*)}.$$

The characteristic good g^* usually corresponds to a central tendency indicator of the characteristics' vector. Characteristic hedonic price indices basically possess the same advantages of imputed hedonic price indices over time-dummy indices. It is not clear if imputed or characteristic price indices provide a better approach for estimating price indices. The main advantage over imputation indices seems to be that the two different price index formulae are unified in a single formula, thus reducing the price index problem. This advantage seems to be compensated by the problem of choosing a good g^* representative of the set G. If the purpose of the hedonic price index is, for example, to provide a sound indicator of the housing market condition, basing a price index on an average housing good may neglect particular dynamics of certain types of housing goods. Little research, however, analyzed characteristic hedonic price indices, and these questions could provide material for further research.

1.3 STATISTICAL PROBLEMS OF HEDONIC METHODS

In this section, the statistical problems related to estimating equation (1) and computing hedonic price indices are reviewed. These problems are of concern depending on the purpose of the hedonic analysis, but often affect the two categories.

1.3.1 *Heterokedasticity and autocorrelation*

When equation (1) is estimated, the presence of heteroskedasticity is thought to be caused by one of the characteristics included in the hedonic regression. This assumption is usually motivated by the large number of variables included in the

initial model, thus avoiding the presence of omitted variables in the model's error term. In equation (3), moreover, the heteroskedasticity problem may be even more serious, since it is legitimate to assume a time dependency of the error term's distribution. In this case, the characteristics and time may be responsible for heteroskedasticity.

Although often present in hedonic regression models, the heteroskedasticity problem has been explicitly addressed by only a few authors. Using only the living area and age as characteristics of single-family houses, Goodman and Thibodeau (1995) identified the age of a house as the main cause leading to the detection of heteroskedasticity. In a subsequent analysis, Goodman and Thibodeau (1997) included additional characteristics and considered the possible influence of housing submarkets in detecting heteroskedasticity. They found that, although they segmented the housing market into submarkets, age-related heteroskedasticity was observed for half of the market segments. Fletcher et al. (2000) supported these findings, but also identified the external area of the property as a possible characteristic causing heteroskedasticity. The heteroskedasticity issue has been explored by Stevenson (2004a), who also identified age as the main source of heteroskedasticity in hedonic price equations at an aggregate level. He found, however, that the definition of housing submarkets seems to eliminate the heteroskedasticity problem for the majority of the submarkets, and argued that a lesser variation of the characteristics at a disaggregate level may be responsible for this phenomenon.

In fact, it is not clear how the age of a house affects its price. As explained by Goodman and Thibodeau (1995), several opposite effects are usually measured by the age variable. Either the house is old enough to produce a vintage effect and increase the house price, or the age of a house decreases the house price due to deterioration. Moreover, the probability of renovations usually increases with the age of the house, countering the house depreciation. Thus, the more the age of a house increases, the more the variance of the error term is supposed to increase, leading to heteroskedasticity.

To address the heteroskedasticity problem, three main approaches have been adopted in the hedonic literature. The first, often used when the aim is to assess the impact of externalities on the house price, is to use Heteroskedasticity-Consistent (HC) estimators of the coefficients' standard errors. This approach allows the usual testing procedures, and establishes if the given

set of characteristics is significant. White and Leefers (2007) and Nelson (2010) represent recent publications using HC estimators in the case of hedonic regressions. The second approach, mainly used when the focus is on predicting prices, consists of directly modeling the heteroskedastic error term as a function of the characteristics, and successively estimates the hedonic model with iterative reweighted regression methods. This modus operandi, in contrast to the previous approach, generally reduces the variance of the predictions. Goodman and Thibodeau (1995), Goodman and Thibodeau (1997), Fletcher et al. (2000), and Stevenson (2004a) used this approach. The third approach is used in both categories of hedonic studies. In the parametric context, heteroskedasticity is often considered to be caused by a bad model specification. Thus, transformation methods based on the paper by Box and Cox (1964) have been applied to dependent and independent variables to assume a normal distribution of the error term, and then proceed with a maximum-likelihood estimation approach (see Maurer et al. (2004)).

In contrast to heteroskedasticity, temporal autocorrelation is usually not a problem when equation (1) and (3) are estimated, since in this case pooled or simply cross-sectional data are used. More relevant than time correlation is the possible spatial correlation between the observations. In the housing context, location is one of the most important characteristics determining the price of a house. If this characteristic is not correctly taken into account in the hedonic equation, a spatial correlation may be present, possibly leading to inefficient and biased estimators. The definition of a spatial correlation structure in the hedonic price equation is mainly intended to avoid these statistical problems. To implement this structure, however, geo-spatial data are necessary, thus increasing the amount of data necessary for estimation. For a review of the studies considering spatial dependency in the hedonic housing context, see Hill (2012).

Importantly, the consequences of the presence of heteroskedasticity and/or autocorrelation are negative for both categories of hedonic studies. In fact, standard testing procedures are not valid, and the model's predictions are inefficient.

1.3.2 *Functional form*

The functional form problem has been addressed in various ways, depending on the aim of the hedonic study. The stud-

ies trying to assess the impact of a specific set of characteristics have focused more on parametric estimation techniques, whereas valuation and prediction studies have used nonparametric methods extensively.

The first category mainly relied on the Box-Cox transformation methods to allow general model specifications, thus maintaining the interpretation of the characteristics in the price-building process. Recently, however, modern nonparametric techniques, although not allowing standard testing procedures, have also been used to determine the impact of specific characteristics on the price of a house. The second category of hedonic studies seems to be the focus of more research than in the past, probably also due to the recent interest in hedonic price indices. Nonparametric methods are widely used in this category, and are often combined with resampling techniques to assess the model's prediction accuracy. Clearly, studies focusing on prediction often use black-box estimation techniques, and determining the role played by a characteristic in the price-building process is not always possible. See Hill (2012) for a review of nonparametric methods applied to housing goods in the hedonic context. Recently, the flexibility of genetic algorithms has prompted researchers in the hedonic domain to apply the neural networks estimation technique to predict prices. See Landajo et al. (2012) for a review of this specific estimation method in the hedonic domain.

A remark on the hedonic model's functional form is necessary at this point. The goal of a prediction model in hedonic studies must be to achieve a good trade-off between in- and out-of-sample prediction accuracy. Focusing exclusively on the in- or out-of-sample prediction accuracy may lead either to overfitting or underfitting the available data. This concept is of paramount importance in the case of imputed hedonic price indices. In fact, imputed prices for the goods observed in one period correspond to out-of-sample predictions in hedonic model estimated in another time period. In-sample and out-of-sample prediction play thus a role in computing hedonic price indeices. Surprisingly, although often possessing hyperparameters that can be chosen to improve out-of-sample prediction accuracy, nonparametric estimation techniques have not been used extensively to compute hedonic price indices. In addition, probably due to the binary nature of the aim driving hedonic research, semiparametric methods have also been rarely used in the hedonic housing domain in general, and, in particular, to compute hedo-

nic price indices. Up to the present, researchers seem to prefer either full parametric or nonparametric estimation techniques, but this approach may be questionable depending on the hedonic price index that must be computed. In this author's opinion, using a semiparametric approach to estimate equation (3) could provide a more flexible functional form, thus solving one of the problems related to time-dummy price indices.

1.3.3 *Variable selection and multicollinearity*

When the purpose of the statistical analysis is to establish if a specific subset of characteristics affects the price-building process, variable selection methods are used. Since no economic model is available to guide the researcher in selecting relevant characteristics, the bottom-up testing approach is impractical, since the initial variables chosen for inclusion in the hedonic regression are completely subjective. A common strategy adopted by researchers and practitioners is thus to introduce all the available characteristics in the hedonic price function and then use a top-down testing approach. Although this approach provides a better alternative than the bottom-up approach, this approach also presents some, often ignored, considerable drawbacks. First, when several subsequent tests, i.e., multiple tests, are performed, the final confidence level does not correspond to the nominal level of each test. Moreover, the tests are generally not independent of each other, thus requiring sophisticated statistical techniques to compute the final level of the multiple tests. Second, when the number of characteristics is large, it is difficult to choose which variables to test without involving subjective judgements of the variables' importance.
The problems that arise when these testing approaches are used are worsened by the presence of multicollinearity among the characteristics. In this case, the variance of the estimated coefficients is usually high, making the variables not significant. Excluding a not significant variable in the selected model when multicollinearity is present may result in a severe model misspecification, and lead to the wrong conclusions concerning the set of characteristics affecting the price. Moreover, the presence of multicollinear characteristics negatively affects the prediction accuracy of hedonic linear models, and may cause instability in the estimation results for other regression methods.
Although multicollinearity is not a major concern for many fields, it naturally occurs when the hedonic approach is used.

In the hedonic context, according to the author's experience, multicollinearity seems to result from the combination of four situations. In the first situation, the variables contained in a subset of characteristics are linearly dependent because of a technological constraint in the production process of the goods. Technology goods typically belong to this first situation. The second case is when consumers' preferences imply a relation between the observed combinations of characteristic bundles. The number of bathrooms per total number of rooms in a house seems, for example, to be dictated more by consumers' preferences than by technological constraints. The third situation appears when two characteristics are proxies for the same unobservable variable. The physical volume and the number of rooms in a house can be thought of as proxies of a 'space' variable, and will probably be highly correlated. Finally, multicollinearity could arise when polynomial and/or interaction terms of the characteristics are considered in the hedonic function. This case occurs when one has a statistical need to obtain a good fit, describe a non-linear behaviour of a characteristic of interest, or take into account the possible heteroskedasticity of the error term. The variable *age*, for example, is typically considered to cause heteroskedasticity and, therefore, age's quadratic and cubic terms are often included in the main regression. It is thus important to stress that, although multicollinearity may not be present among the original set of characteristics, it could be induced.
Surprisingly, the multicollinearity problem is often neglected in the hedonic literature. Multicollinearity indicators, such as the variance inflation factor and the condition number, are rarely reported in hedonic publications, thus casting shadow on the validity of the selection procedure and on the predictions' accuracy.

1.3.4 *Hedonic indices and the price index problem*

In the present context of heterogeneous goods, the price index problem is defined as the inability to choose a specific price index formula in (2). This choice exists not only in selecting a general price index formula but also in adopting a central tendency indicator μ. Two main approaches have been suggested in the price index literature to address the price index problem. The first approach, based on the economic theory of utility and cost functions, demonstrates how each price index formula cor-

responds to specific assumptions made about the utility or cost functions. Depending on the validity of these assumptions, a price index can thus be selected. The second approach uses the axiomatic theory developed in the price index literature. This approach assumes that a price index must satisfy a certain number of axioms to effectively measure price changes. To choose among different price index formulae, a given set of axioms is chosen. Each axiom is then checked for each index in turn.
These approaches, however, were developed for classic price indices. Beer (2006) adopted an axiomatic approach to verify the axiomatic properties of general hedonic elementary price indices. To the author's knowledge, Beer's work represents the first attempt to solve the price index problem in the hedonic context. More recently, Hill and Melser (2008) investigated the hedonic price index problem for housing goods. They concluded that the hedonic approach complicates the index problem, and introduced a new source of variation in the indices. Up to the present, the price index problem has been neglected in the hedonic literature, and requires further research to be addressed.

1.4 PERSONAL CONTRIBUTION

In light of the statistical problems, four research papers have addressed one specific problem in turn. The aim was to extend, through an empirical or theoretical approach, actual knowledge in the hedonic field, providing results that are useful for researchers and practitioners.
The price index problem in the hedonic context is discussed in the first two papers. The third paper focuses on variable selection in hedonic models where multicollinearity is present. In particular, a new variable selection algorithm that outperforms ordinary automated selection techniques is proposed.
The fourth paper implements a methodology for comparing the prediction accuracy of two hedonic models. The empirical results of these papers are all based on a data set kindly provided by Wüest & Partner, an international real estate consultancy firm. Transaction prices for single-family dwellings and their corresponding characteristics were collected for the Swiss canton of Zurich from banks, insurance companies, and other real estate agencies. The collected data are organized in 44 subsequent quarterly data sets, spanning the first quarter of 2001 to the fourth quarter of 2011.

1.4.1 *The econometric foundations of hedonic elementary price indices*

The first article is a paper originally written by Brachinger and Beer (2009), thoroughly revised and extended by Dr Michael Beer and the author in 2012 (see Brachinger et al. (2012)). It proposes a mathematical description of characteristics and elementary aggregates. In the following step, a hedonic econometric model is formulated, and hedonic elementary population indices based on (2) are explicitly defined by choosing a central tendency indicator μ. We emphasise that population indices are unobservable economic parameters that must be estimated with suitable sample indices. It is shown that, within the developed framework, many of the hedonic index formulae used in practice are identified as sample versions corresponding to particular hedonic elementary population indices.
Using the introduced theoretical framework, a general procedure for estimating the confidence intervals of hedonic elementary price indices is then proposed. This procedure is implemented in the empirical part of the paper - the author's principal contribution -, where the hedonic indices are estimated along with their bootstrapped confidence intervals. To compare the sample variation of the hedonic price indices, the confidence intervals' lengths are then computed. The obtained confidence intervals' lengths, together with the results from price index theory, suggest an empirical answer to the price index problem. This conclusion partially sheds light on the price index problem, allowing practitioners to choose among different price index formulae.

1.4.2 *Asymptotic properties of imputed hedonic price indices in the case of linear hedonic functions*

In the second paper, the asymptotic properties of the most common hedonic price indices based on the elementary price indices presented in (2) have been analyzed. In particular, the hedonic counterpart of the Laspeyres, Paasche, and Fisher price index has been considered for the single, double, and characteristic methods. In fact, although these indices are used widely, little research has been conducted to investigate the asymptotic properties of these hedonic price indices. The present paper therefore attempts to fill the actual knowledge gap by analyzing the asymptotic properties of the most commonly used im-

puted hedonic price indices.
The analysis must be considered as a new theoretical approach for addressing the price index problem, relying on the stochastic component introduced by the hedonic approach, and usually not present in ordinary price indices. Interestingly, the asymptotic equivalence of single imputed, double imputed, and characteristics hedonic price indices has been established in the case of a linear hedonic function. This result appears to be quite important, since it implies that the price index problem tends to vanish in large samples from a probabilistic point of view, thus alleviating an uncomfortable situation price statisticians have to face.
Despite the importance of the results, they must also be carefully placed in the context in which estimating hedonic models takes place. In the case of a non-linear hedonic regression function, our results are generally not valid, even for continuous hedonic functions. An important case is represented by log-linear hedonic models, which are commonly used to model the hedonic prices of housing goods.
A secondary result of the paper concerns the asymptotic distribution of hedonic price indices. Due to the high nonlinearity of the parameters present in the hedonic price index formulae, it seems unrealistic to analytically compute the asymptotic distribution of such indices even in the case of linear hedonic function, which suggests the resampling methods described in the first paper should be used.

1.4.3 *A new approach to variable selection in the presence of multicollinearity: a simulated study with hedonic housing data*

The third paper evaluates the impact of multicollinearity on automated variable selection procedures. In particular, backward stepwise selection based on the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) criteria is considered. This objective is achieved by comparing the model selection performance of various selection methods in hedonic regression models where noise variables inducing or not inducing multicollinearity have been introduced. In addition to analyzing widespread stepwise selection methods based on information criteria, a new selection method using a multimodel approach is also examined.
The performance of the selection procedures is gauged regarding their ability to identify the data-generating process. To this

end, a data-generating process is simulated. The paper exclusively focuses on identifying this data-generating process, assuming that the process is a nested version of a more general model in which all the available characteristics are included in the analysis. The ability of each selection method to correctly identify this data-generating process is compared for different simulation settings.
The findings suggest that, when multicollinearity is not present, backward stepwise regression is a reliable method for identifying the original data-generating process. The performance of this selection method, however, is decreased in the case of noise variables inducing multicollinearity, thus requiring the use of an alternative selection approach. To solve this problem, a selection method based on a multimodel approach has been suggested. The proposed method clearly shows better performance in the case of multicollinearity and, surprisingly, seems to perform equally well with the AIC and BIC criteria. These two features are particularly important, since even if multicollinearity is present and the non-appropriate information criterion for the set-up is used, the correct set of informative variables is in general selected.

1.4.4 *Selection of regression methods in hedonic price models based on prediction loss functions*

The fourth paper analyzes how the selection of a regression method in hedonic price models is affected by the loss function used to assess the model's prediction accuracy. To this aim, a given hedonic price model is estimated using an ordinary least squares regression method and a robust regression technique. The predictive accuracy of each regression method is computed for various loss functions, and then compared to the prediction accuracy of the rival estimation method. This comparison is performed by difference in means t-tests on prediction losses.
Up to the present, only a few studies have illustrated the advantages of robust estimation methods in hedonic models when the main goal is to analyse characteristics' implicit prices. It is not clear, however, whether estimation techniques perform better than usual OLS estimators from a prediction point of view, and whether this performance depends on the loss function used to compute the prediction error.
To assess the models' prediction accuracy, losses caused by predictions obtained with a classic linear regression model are

compared to losses caused by predictions obtained with robust estimation techniques. This intent is achieved by carrying out mean-difference tests for paired samples on the prediction losses. In particular, three loss functions were considered in the present analysis: square, absolute, and bisquare loss function.
A rough comparison of the expected loss point estimates, as often conducted in the hedonic literature, is shown to be misleading, since it does not account for the sample variation of these point estimates. Moreover, the out-of-sample predictive accuracy may depend strongly on the resampling technique used to compute out-of-sample predictions. Great care is thus needed in interpreting the out-of-sample prediction accuracy. Based on the methodology, the robust estimator was shown to perform as well as the OLS estimator for the square and absolute loss functions. Interestingly, although the two regression methods showed similar coefficients' values, the robust estimator performed significantly better for the bisquare loss function.

Part II

RESEARCH PAPERS

2

THE ECONOMETRIC FOUNDATIONS OF HEDONIC ELEMENTARY PRICE INDICES [1]

2.1 INTRODUCTION

A Consumer Price Indices (CPI) measures the average price change of consumer goods in a market between two fixed time periods, assuming that their quality remains constant. In practice, however, the quality of the universe of products that households consume is continually changing. It is therefore necessary to estimate the contribution of the quality change to the observed price change in order to measure the quality-corrected 'pure' price change.

The state-of-the-art manner of handling differences and changes in quality is the so-called *hedonic approach*. Its main idea is to identify the quality of a product – or, in other words, its 'potential contribution ... to the welfare and happiness of its purchasers and the community' (Court, 1939, p. 107) – with a vector of product characteristics. In the hedonic approach, a regression equation is estimated relating the characteristics of the product to its price. Once such a relationship is established, the price of any similar item can be predicted by plugging its characteristics into the estimated *hedonic (regression) function*.

CPI concepts usually structure the basket of consumer goods in a hierarchical way. Individual price observations are transformed into a final index value through a sequence of aggregation steps. In the first stage, the price evolution is individually observed for restricted groups of homogeneous products, the so-called *elementary (expenditure) aggregates*. These aggregates usually serve as strata for data collection and form the building blocks of a CPI. For each of them, a so-called *elementary price index* is calculated. In further stages, these elementary price indices are 'averaged to obtain higher-level indices using the relative values of the elementary expenditure aggregates as weights' (ILO et al., 2010, para. 9.3). The need for adjusting price measurements for quality change normally appears at the level of elementary aggregates when individual prices are directly com-

1 This article is a joint work by Prof. Hans Wolfgang Brachinger, Dr Michael Beer and the author (see Brachinger et al. (2012)).

pared. Therefore, quality adjustment is – to our opinion – predominantly an issue of elementary price indices. Elementary price indices where the quality adjustments are based on the hedonic approach are called *hedonic elementary price indices*.

The literature on hedonic methods in price statistics is steadily growing, with Triplett (2004), ILO et al. (2010, Chap. 21) and Hill (2011) being three of the most recent comprehensive and fundamental overviews – the last of them with a particular focus on house price indices. The present paper contributes to this literature in proposing a formal framework for hedonic elementary price indices that incorporates and generalises these approaches. Our framework corroborates the existing theory by providing a novel conceptual approach from which most of the elementary hedonic price indices used in practice can be derived. Moreover, it defines the necessary concepts that allow, e.g., to examine state-of-the-art hedonic index estimators from both an axiomatic and empirical viewpoint. We emphasise particularly the clear separation of elementary (population) indices as unobservable economic parameters from their estimators, the sample indices. In this aspect, the present piece of work abuts on the mindset of papers like, e.g., Dorfman et al. (1999), Brachinger (2002), Balk (2005), and Silver and Heravi (2007).

Section 2.2 provides a precise definition of characteristics and elementary aggregates. This definition permits a clear-cut and concise formulation of an econometric model underlying every hedonic price index. Section 2.3 discusses elementary price index concepts in general. These concepts are extended to universal formulae for hedonic elementary population indices in Section 2.4. Each of these indices is a well-defined economic parameter that eventually needs to be estimated from a random sample of observations. In Section 2.5 sample versions of the universal formulae are presented and a general procedure to estimate their confidence intervals is proposed. It is shown that from these sample indices most of the elementary indices used in practice can be derived. An empirical application of the introduced concepts to residential housing prices is finally presented in section Section 2.6. The paper closes with a short summary.

2.2 ELEMENTARY AGGREGATES AND THE HEDONIC ECONOMETRIC MODEL

2.2.1 *Goods and characteristics*

We begin by describing the basic entities of an elementary price index, namely some consumer goods offered in a market, the set of characteristics they exhibit, and the corresponding elementary aggregate. The formal language used for this purpose will allow us later to build an econometric model on top. From the outset, the characteristics of the goods play an important role as surrogates for the notion of quality. Their omnipresence in our framework will make the step from general (quality-unadjusted) to hedonic (quality-adjusted) elementary price indices straightforward.

Let $\mathcal{O}$ denote the set of all consumer goods supplied in a market at a certain point in time.[2] Here, the notion of a *good* means physically tangible items like, e.g., used cars or personal computers as well as services and other immaterial entities to which a price can be assigned. Each of these goods exhibits a set of *characteristics*. Examples of such characteristics might be the volume or the physical mass of the good, the horsepower, mileage or colour of a used car, or the processor speed of a computer. Other non-physical characteristics comprise the location of sale or any after-sales service. Statistically speaking, a characteristic simply is a variable or an attribute. It may be scaled on different measurement levels from nominal to cardinal.

It is obvious that not every characteristic can be observed for a given good $o \in \mathcal{O}$, i.e. each characteristic m is generally only defined on a specific subset $\mathcal{O}_m$ of $\mathcal{O}$. Processor speed, for example, is a characteristic of computers but not (yet) of clothes or bicycles. So the domain $\mathcal{O}_m$ of the characteristic m : 'processor speed' contains the set of all computers, but also all other devices carrying a CPU including many modern household appliances, cars, and communication devices. m might not always be relevant to the purchaser of such goods, but technically, it is defined and measurable; we will come back to the economic relevance of certain characteristics later. Food, hospital services, and package holidays are examples of goods for which the characteristic 'processor speed' is not defined, hence they lie outside $\mathcal{O}_m$.

2 We are going to raise the restriction to a single point in time in Section 2.4.

Although characteristics can be of any measurement scale, it is always possible to quantify their values such that they form a subset of the Euclidean real number space. This leads to the following definition:

Definition 2.2.1. *A **characteristic** m is a real-valued function $m : \mathcal{O}_m \longrightarrow \mathbb{R}$ defined on a non-empty subset $\mathcal{O}_m$ of $\mathcal{O}$. The set $\mathcal{O}_m$ is called the **domain** of m and, for each $o \in \mathcal{O}_m$, $m(o)$ will be called the m-**value** of o.*

For the sake of simplicity, the m-value of o may also be called its m-characteristic. The set of all characteristics will be denoted by $\mathcal{M} := \{m : \mathcal{O}_m \longrightarrow \mathbb{R} \,|\, \mathcal{O}_m \subset \mathcal{O}, \mathcal{O}_m \neq \emptyset\}$.

The reason why we put such emphasis on the domains of the characteristics is that they will now serve as building blocks for our definition of an elementary aggregate. Guidelines to practitioners on how elementary aggregates should be specified are traditionally rather vague and leave most decisions to the users' discretion. The authors of ILO et al. (2010, para. 9.7), e.g., confine themselves to requiring that elementary aggregates consist of goods that are 'as similar as possible', 'preferably fairly homogeneous', 'expected to have similar price movements', and 'appropriate to serve as strata for sampling purposes'. While such formulations may suffice in practice, they are much too cursory to serve as a building block of a hedonic econometric model. The following much more formal definition of an elementary aggregate contains all elements needed for the elaboration of our econometric framework.

Definition 2.2.2. *An **elementary aggregate** $\mathcal{G}$ is a set of goods in $\mathcal{O}$ having the following properties:*

1. *The set $\mathcal{M}_\mathcal{G}$ of the characteristics defined for all elements of $\mathcal{G}$ is not empty, i.e.*

$$\mathcal{M}_\mathcal{G} := \{m \in \mathcal{M} \,|\, \mathcal{O}_m \supset \mathcal{G}\} \neq \emptyset\,. \tag{6}$$

2. *The intersection of the domains of all characteristics contained in $\mathcal{M}_\mathcal{G}$ is a subset of $\mathcal{G}$, i.e.*

$$\bigcap_{m \in \mathcal{M}_\mathcal{G}} \mathcal{O}_m \subset \mathcal{G}\,. \tag{7}$$

*Each element $o \in \mathcal{G}$ will be called an **item** of the elementary aggregate $\mathcal{G}$. The elements of $\mathcal{M}_\mathcal{G}$ are called **distinguishing characteristics** of $\mathcal{G}$.*

Property (1) means that all goods belonging to an elementary aggregate $\mathcal{G}$ have at least one characteristic in common. Conversely, if two goods do *not* belong to the same elementary aggregate, there must be a characteristic that is defined for one of these goods but not for the other. Property (2) ensures that each good carrying all characteristics of $\mathcal{M}_\mathcal{G}$ is contained in the elementary aggregate. Note that every elementary aggregate in the sense of Def. 2.2.2 is defined relative to the set $\mathcal{O}$ of all goods supplied on the market.

The following proposition shows that each elementary aggregate has some kind of maximality property in the sense that its distinguishing characteristics fully determine the items of the aggregate. In other words, there is no item of an elementary aggregate that is not contained in the intersection of the domains of all distinguishing characteristics.

Theorem 2.2.1. *Each elementary aggregate $\mathcal{G}$ equals the intersection of the domains of its distinguishing characteristics, i.e.*

$$\bigcap_{m \in \mathcal{M}_\mathcal{G}} \mathcal{O}_m = \mathcal{G}\,. \tag{8}$$

Proof 2.2.1. *We have $\mathcal{G} \subset \mathcal{O}_m$ for all $m \in \mathcal{M}_\mathcal{G}$. Therefore, $\mathcal{G} \subset \bigcap_{m\in\mathcal{M}_\mathcal{G}} \mathcal{O}_m$. The inclusion in the other direction is given by property 2 of Def. 2.2.2, hence equality holds.*

It should be noted that an elementary aggregate in the sense of Def. 2.2.2 may still comprise many different items. In particular, there is no explicit requirement regarding the similarity or homogeneity of the items contained. So if, e.g., the physical mass of an object was taken as the only distinguishing characteristic, the respective elementary aggregate would embrace the whole universe of physically tangible goods, excluding just services and other intangible products like computer software. Thus we define the term 'elementary aggregate' in a much broader sense than it is usually applied in practice. However, it follows from Prop. 2.2.1 that supplementing the set $\mathcal{M}_\mathcal{G}$ of distinguishing characteristics of an elementary aggregate with additional characteristics leads to a diminution of $\mathcal{G}$. By selecting the appropriate list of distinguishing characteristics, we may thus in practice reduce a very general aggregate to one which satisfies the homogeneity or similarity requirements cited above.

As a consequence of Prop. 2.2.1, it is furthermore possible to induce elementary aggregates from samples of individual

goods. Let $\mathcal{O}^* \subset \mathcal{O}$ be any set of goods. These might be, e.g., different models of personal computers. Let $\mathcal{M}_{\mathcal{O}^*} := \{m \in \mathcal{M} \,|\, \mathcal{O}_m \supset \mathcal{O}^*\}$ be the set of all characteristics whose domains contain these goods, i.e. all characteristics that are defined for all elements of $\mathcal{O}^*$. In the case of personal computers, these would contain typical attributes such as CPU speed, RAM size, hard drive size, brand, length of warranty period, etc., but also others such as the serial number, which may not be relevant to the consumers' purchase decision. Assume that $\mathcal{M}_{\mathcal{O}^*}$ is not empty. Then, it is possible to specify the elementary aggregate $\mathcal{G}(\mathcal{O}^*)$ induced by $\mathcal{O}^*$. The *induced elementary aggregate* is defined as the intersection of the domains of all characteristics in $\mathcal{M}_{\mathcal{O}^*}$, i.e.

$$\mathcal{G}(\mathcal{O}^*) := \bigcap_{m \in \mathcal{M}_{\mathcal{O}^*}} \mathcal{O}_m \,. \tag{9}$$

The set $\mathcal{O}^*$ is thus extended by all goods on the market that exhibit at least the same characteristics as the goods fixed in $\mathcal{O}^*$. Obviously, by means of (9), any given set of characteristics $\mathcal{M}$ induces an elementary aggregate $\mathcal{G}(\mathcal{M}) := \bigcap_{m \in \mathcal{M}} \mathcal{O}_m$.

Def. 2.2.2 of an elementary aggregate is admittedly guided by theoretical elegance rather than practical pertinence. Nobody will be able to provide a comprehensive list of the distinguishing characteristics of even the simplest elementary aggregate being used in practice. Nonetheless is such an abstract definition inevitable to make the vague notion of an elementary aggregate manageable from an econometric viewpoint. Moreover, we presume that the idea of distinguishing characteristics can serve as an implicit guideline for practitioners needing to decide on which items should belong to a certain elementary aggregate. The authors of ILO et al. (2004, para. 3.147 ff.) identify three main approaches to the classification of consumer goods, namely the classification by product type, by purpose, and by economic environment. Recommended practice is 'to use a purpose classification at the highest level, with product breakdowns below'. Inevitably, the characteristics of goods play a certain role when elementary aggregates are defined by product type at the lowest level. The main merit of the concept introduced in Def. 2.2.2 is thus the duality between the elementary aggregate and its distinguishing characteristics. This duality will be exploited below.

As a final note to this first section, it is worth highlighting that the distinguishing characteristics of an elementary aggregate provide some very useful means of identifying items that

are 'equivalent' in a certain sense. We will use this property later to partition an elementary aggregate into classes of equivalent quality. Let $\{m_1, \ldots, m_K\} \subset \mathcal{M}_\mathcal{G}$ denote any finite subset of the distinguishing characteristics of an elementary aggregate $\mathcal{G}$. By assembling them to a vector function

$$\mathbf{m} : \mathcal{G} \longrightarrow \mathbb{R}^K, \quad o \longmapsto \mathbf{m}(o) := (m_1(o), \ldots, m_K(o))' \tag{10}$$

all the items of $\mathcal{G}$ are mapped to a K-dimensional vector of characteristics. This identification of goods with a characteristics vector leads to an equivalence relation on $\mathcal{G}$ defined by

$$o_1 \sim_\mathbf{m} o_2 \quad :\Longleftrightarrow \quad \mathbf{m}(o_1) = \mathbf{m}(o_2)\,. \tag{11}$$

Two items of an elementary aggregate are thus identified if and only if their $\mathbf{m}$-values, i.e. their m_1- to m_K-values coincide. The equivalence classes respective to the relation $\sim_\mathbf{m}$ will be called $\mathbf{m}$-equivalence classes. They partition $\mathcal{G}$ into subsets containing items with equal $\mathbf{m}$-values. The quotient set induced by this equivalence relation will be denoted by $\mathcal{G}/\sim_\mathbf{m}$.

2.2.2 *Characteristics and prices*

In the last section, we identified a good with a list of characteristics and we showed how goods can be grouped into elementary aggregates. The economic foundation of this approach is the consumer theory developed by Lancaster (1966, 1971). This theory assumes that 'one demands not just physical objects, but the qualities with which they are endowed' (Milgate, 1987, p. 546). Consumers' preferences are therefore originally directed towards the characteristics of a good, and the latter determine eventually the consumers' preference ordering between individual items of an elementary aggregate.

Lancaster (1971, p. 140 ff.) himself emphasised that some characteristics of a good are usually irrelevant for a consumer's purchase decision (such as the serial number of a personal computer). Irrelevant characteristics are especially those that are invariant for all items of an elementary aggregate. Inversely, Lancaster defines a characteristic as *relevant* when ignoring it would change the preference ordering between two items.

Driven by the consumers' individual preferences, Lancaster's approach suggests that a good's price observed on the market is essentially determined by the relevant characteristics of that good. This assumption is called *hedonic hypothesis* in the literature (see e.g. Triplett, 1987; United Nations, 1993; Dickie et al.,

1997). The hedonic hypothesis serves as the general basis for all hedonic price indices. In order to build up a solid theory of hedonic price indices, we propose formulating it as an econometric model in the following form:

HEM 2.2.1. *Let $\mathcal{G}$ be an elementary aggregate with distinguishing characteristics $\mathcal{M}_\mathcal{G}$. There exists a finite set of characteristics*

$$\mathcal{M}_\mathcal{G}^{pr} = \{m_1, \ldots, m_{K_\mathcal{G}}\} \subset \mathcal{M}_\mathcal{G} \tag{12}$$

and a function $h_\mathcal{G} : \mathbb{R}^{K_\mathcal{G}} \longrightarrow \mathbb{R}_{\geqslant 0}$, such that the price $p(o)$ of any item $o \in \mathcal{G}$ can be written as

$$p(o) = h_\mathcal{G}\big(\mathbf{m}_\mathcal{G}^{pr}(o)\big) + \epsilon(o) \tag{13}$$

with $\mathbf{m}_\mathcal{G}^{pr}(o) = (m_1(o), \ldots, m_{K_\mathcal{G}}(o))'$. The residual term $\epsilon(o)$ is assumed to be stochastic with conditional expectation

$$E\big(\epsilon(o) \mid \mathbf{m}_\mathcal{G}^{pr}(o)\big) = 0\,. \tag{14}$$

*The set $\mathcal{M}_\mathcal{G}^{pr}$ will be called the **set of price-relevant characteristics**, and $h_\mathcal{G}$ is the **hedonic function** of $\mathcal{G}$.*

This model exploits the idea that for an elementary aggregate for which the hedonic hypothesis holds, the set of distinguishing characteristics contains a finite subset of price-relevant characteristics. They determine the price up to a residual term that covers any quality-independent price component. Assumption (14) implies that the hedonic price of an item with a certain quality is given by the average price over all items of the same quality.

One of the central points here is that the vector of price-relevant characteristics is seen as a surrogate for a good's *quality*. Much in the spirit of the outline provided in the *System of National Accounts 1993* (see United Nations, 1993, para. 16.105 ff.), the term 'quality' subsumes all characteristics of an item which make it distinguishable from other items from an economic point of view. The hedonic function $h_\mathcal{G}$, being defined on the quotient set $\mathcal{G}/\sim_{\mathbf{m}_\mathcal{G}^{pr}}$, maps each class of items of equivalent quality to a constant price.

Assumption (14) appears reasonable if all the equivalence classes $[o]$ are sufficiently homogeneous, which is the case when the number of price-relevant characteristics is large enough. Similarly to what was already mentioned above, the size and thus the homogeneity of the individual classes is a non-increasing function of the number of price-relevant characteristics, since

adding additional characteristics generally leads to more and thus smaller equivalence classes.

We deliberately do not impose any restrictions on the functional form of the hedonic function since, at this stage, we see no reason to do so. Finding an appropriate candidate of a hedonic function that links the vector of price-relevant characteristics to the average price a consumer needs to pay for an item of equivalent quality is a purely statistical issue. Triplett (2004) convincingly argues that 'imposing some rule for what the hedonic function "should" look like destroys part of the information that market prices convey'. Referring to Rosen (1974), he emphasises that 'the form of the hedonic function is entirely an empirical matter that is determined by the distributions of buyers around the hedonic surface, and not by the form of their utility functions.' Therefore, the hedonic function can neither provide an economic explanation for the behaviour of economic agents nor identify demand or supply. It just describes the statistical relationship between the market price and the quality of a good, no matter how the price and thus the purchasers' valuation of characteristics emerge.

In the framework developed so far, we described the universe of consumer goods available in a market at a certain point in time. By means of distinguishing and price-relevant characteristics, we provided a formal definition of the notion of an elementary aggregate and established a link between the price of a good and its quality. The following section now introduces the time dimension and defines the basic forms of elementary price indices.

2.3 ELEMENTARY PRICE INDICES

2.3.1 *Elementary aggregates over time*

Elementary price indices measure the average price evolution of an elementary aggregate between two time periods. There is a base period 0, serving as reference period, and a current period 1 for which the prices are compared. As time passes, we may observe a certain variation of the items contained in a given elementary aggregate: new items appear on the market and are purchased by consumers, others disappear. This effect is particulary pronounced for products where there is a rapid turnover of differentiated models, such as computers, communication, and multimedia devices.

The duality between elementary aggregates and distinguishing characteristics introduced in the previous section allows for a constant understanding of the nature of an elementary aggregate over time. Instead of fixing the exact content of an aggregate, we fix the distinguishing characteristics and allow new objects to become part of the aggregate if and only if they carry at least all these fixed characteristics. More formally, this leads to the following definition of a current (elementary) aggregate.

Definition 2.3.1. *Let $\mathcal{T}$ be the set of all time periods considered and let, for any time $t \in \mathcal{T}$, denote $\mathcal{O}^t$ the set of all goods supplied on the market at time t. Let $\mathcal{G} = \mathcal{G}^0$ be any elementary aggregate defined relative to $\mathcal{O}^0$ and let $\mathcal{M}_\mathcal{G}$ be its set of distinguishing characteristics.*

*Then, for any time $t \in \mathcal{T}$, the **current aggregate** $\mathcal{G}^t = \mathcal{G}^t(\mathcal{M}_\mathcal{G})$ is defined as the elementary aggregate induced by $\mathcal{M}_\mathcal{G}$ on $\mathcal{O}^t$, i.e.*

$$\mathcal{G}^t(\mathcal{M}_\mathcal{G}) := \bigcap_{m \in \mathcal{M}_\mathcal{G}} \mathcal{O}^t_m , \tag{15}$$

where $\mathcal{O}^t_m \subset \mathcal{O}^t$ denotes the domain of characteristic m in period t.

*The **composite elementary aggregate** $\mathcal{G}^\mathcal{T}$ **induced by** $\mathcal{G}^0$ is defined by*

$$\mathcal{G}^\mathcal{T} := \bigcup_{t \in \mathcal{T}} \mathcal{G}^t(\mathcal{M}_\mathcal{G}) . \tag{16}$$

Obviously, by means of (16), any elementary aggregate $\mathcal{G}$ defined relative to a base period induces a composite elementary aggregate $\mathcal{G}^\mathcal{T}$ for any set $\mathcal{T}$ of time periods. Technically, $\mathcal{G}^t(\mathcal{M}_\mathcal{G})$ can be empty for certain $t \in \mathcal{T}$.

If we focus on the bilateral comparison of a reference period 1 with a base period 0, one feature of our approach is that it yields with $\mathcal{G}^0 \cap \mathcal{G}^1 \subset \mathcal{G}^{\{0,1\}}$ a straightforward identification of the set of matched items for the two periods. Moreover, the disappearing items are assembled in the difference set $\mathcal{G}^0 \setminus \mathcal{G}^1$ while any new unmatched items are represented by $\mathcal{G}^1 \setminus \mathcal{G}^0$. The impossibility of matching price observations over time being *the* main motivation for applying quality adjustment techniques in price statistics, these difference sets are going to be of particular importance in our framework.

2.3.2 *Concepts of elementary price indices*

Relating to what was said in the last paragraph, an elementary price index is typically calculated from two sets of matched

price observations: individual goods are sampled from an elementary aggregate and their prices are collected over a succession of time periods. For the bilateral comparison of two time periods 0 and 1 this implies that only those items of the elementary aggregate $\mathcal{G} = \mathcal{G}^0$ are considered which remain available in period 1. Moreover, items newly appearing in period 1 are ignored as their price cannot be matched with a price in the base period. In other words, bilateral comparisons are *a priori* restricted to $\mathcal{G}^0 \cap \mathcal{G}^1$ (and raising this restriction will be the central purpose of quality-adjusted price indices).

There are basically two competing approaches for the specification of an elementary price index. One approach relates the *average price* of the elementary aggregate $\mathcal{G}$ in the current period 1 to its average price in the base period 0, whereas the other takes the *average price ratio* of the individual items as a measure for the change in price level observed from 0 to 1.

If we denote by μ a measure of location defined for any univariate distribution of positive real numbers (i.e. what we called 'average' above), the two approaches just described can be written as

$$EPI^{0:1}(\mathcal{G}) = \frac{\mu\big(\tilde{p}^1(\mathcal{G})\big)}{\mu\big(\tilde{p}^0(\mathcal{G})\big)} \tag{17}$$

and

$$EPI^{0:1}(\mathcal{G}) = \mu\left(\widetilde{p^1/p^0}(\mathcal{G})\right), \tag{18}$$

respectively.[3] Note that these indices are population indices since they are defined on the whole population of items of a given elementary aggregate. As such, they are latent economic parameters that cannot be observed in practice.

Several elementary price index formulae co-exist in statistical practice which must be considered as functions 'that transform sample survey data into an index number' (Balk, 2005, p. 676) or, in other words, as an estimator or the sample version of a population index.[4] They base upon a sample of objects $o_1, \ldots, o_N \in \mathcal{G}^0 \cap \mathcal{G}^1$ available in both periods for which prices $p_n^t := p^t(o_n)$ were collected. The most widely used formulae

3 In these formulae, $\tilde{p}^t(\mathcal{G})$ stands for the distribution of the prices $\{p^t(o)\}$ and $\widetilde{p^1/p^0}(\mathcal{G})$ for the distribution of the price ratios $\left\{p^1(o)/p^0(o)\right\}$ of all items $o \in \mathcal{G}^0 \cap \mathcal{G}^1$ with $p^t(o)$ being the observed price of an item o at time t ($t \in \{0, 1\}$).

4 We further refer to the papers by Dorfman et al. (1999) as well as Silver and Heravi (2007) for some discussion on the fundamental distinction between sample and population indices.

Population index	Sample index	Index type
(17)	$\widehat{EPI}_{D}^{0:1} = \frac{\sum_{n=1}^{N} p_n^1}{\sum_{n=1}^{N} p_n^0}$	Dutot
	$\widehat{EPI}_{J}^{0:1} = \frac{\sqrt[N]{\prod_{n=1}^{N} p_n^1}}{\sqrt[N]{\prod_{n=1}^{N} p_n^0}}$	Jevons
	$\widehat{EPI}_{HD}^{0:1} = \frac{\left(\sum_{n=1}^{N} (p_n^1)^{-1}\right)^{-1}}{\left(\sum_{n=1}^{N} (p_n^0)^{-1}\right)^{-1}}$	'Harmonic Dutot'
(18)	$\widehat{EPI}_{C}^{0:1} = \frac{1}{N} \sum_{n=1}^{N} \frac{p_n^1}{p_n^0}$	Carli
	$\widehat{EPI}_{J}^{0:1} = \sqrt[N]{\prod_{n=1}^{N} \frac{p_n^1}{p_n^0}}$	Jevons
	$\widehat{EPI}_{HC}^{0:1} = \left(\frac{1}{N} \sum_{n=1}^{N} \left(\frac{p_n^1}{p_n^0}\right)^{-1}\right)^{-1}$	'Harmonic Carli'

Table 1: Elementary sample indices.

are summarised in Table 1. They differ in the population index they target and in the way they implement the measure of location μ, namely, e.g., as arithmetic, geometric, or harmonic mean. The Jevons elementary price index formula targets both population indices simultaneously, since (17) and (18) coincide if μ is implemented by the geometric mean.

There has been much debate in the literature on which of these and other alternative elementary sample indices was the most favourable. We do not intend to take this discussion any further but refer to Chapter 20 of ILO et al. (2004) for a detailed and comprehensive overview. It is just worth highlighting that the discussion on what index type to prefer should start at the level of population indices where no sampling issues arise. If there is no apparent economic reason to favour either (17) or (18) and any specific choice of μ, there may be axiomatic and empirical arguments that lead to a preferred definition. We are going to take up this point later when we look at hedonic elementary population indices.

The most important issue of the elementary price indices introduced so far is their inability to cope with a changing universe of items contained in the elementary aggregate. Limiting the set of items to those which are available in all time periods considered is, in general, a far too restrictive strategy. For many specific aggregates, especially for those subject to rapid technological progress, the set of items available in the base period *and* in all current periods will be too small to represent well enough the range of items of the aggregate.

Therefore, the set of items for which prices are available in all time periods considered must be artificially enlarged. This is usually done by assigning ('imputing') estimated prices to those items of the aggregate which are unobservable in certain time periods. Conventional methods for imputing unobserved prices are typically *ad hoc* solutions that attempt to deduce the price of an item by 'quality-adjusting' the observed price of another item of similar quality (see e.g. ILO et al., 2004, Chap. 7 or Triplett, 2004, Chap. II for a comprehensive overview). However, they lack a sound methodological foundation and may not work consistently for all individual items of an elementary aggregate. A more satisfying solution to the problem of imputing unobservable prices is offered by the hedonic approach.

Based on the hedonic econometric model introduced in Section 2.2, we are now going to extend the two population indices outlined above such that they incorporate the entire population of a composite elementary aggregate.

2.4 HEDONIC ELEMENTARY PRICE INDICES

2.4.1 *The hedonic econometric model revisited*

The hedonic econometric model establishes a relationship between characteristics and prices of the items of an elementary aggregate. This relationship is valid for a fixed point in time. In order to explicit its time dependency and to facilitate the comparison of two or more time periods, we propose to reformulate the model as follows:

HEM 2.4.1 (in time)**.** *Let $\mathcal{G} = \mathcal{G}^0$ be any elementary aggregate defined relative to the set $\mathcal{O}^0$ of all goods supplied on the market at a base period 0 and let $\mathcal{M}_\mathcal{G}$ be its set of distinguishing characteristics. Let $\mathcal{G}^\mathcal{T}$*

be the composite elementary aggregate induced by $\mathcal{G}^0$ *for a given set of time periods* $\mathcal{T}$*. There exists a finite set of characteristics*

$$\mathcal{M}_{\mathcal{G}}^{\mathrm{pr}} = \{m_1, \ldots, m_{K_{\mathcal{G}}}\} \subset \mathcal{M}_{\mathcal{G}} \tag{19}$$

and, for each period $t \in \mathcal{T}$*, a function* $h_{\mathcal{G}}^t : \mathbb{R}^{K_{\mathcal{G}}} \longrightarrow \mathbb{R}_{\geqslant 0}$*, such that the price* $p^t(o)$ *of any item* $o \in \mathcal{G}^t$ *available at time* t *can be written as*

$$p^t(o) = h_{\mathcal{G}}^t\big(\mathbf{m}_{\mathcal{G}}^{\mathrm{pr}}(o)\big) + \epsilon^t(o) \tag{20}$$

with $\mathbf{m}_{\mathcal{G}}^{\mathrm{pr}}(o) = (m_1(o), \ldots, m_{K_{\mathcal{G}}}(o))'$*. For all* $t \in \mathcal{T}$*, the residual term* $\epsilon^t(o)$ *is assumed to be stochastic with conditional expectation*

$$\mathrm{E}\big(\epsilon^t(o) \mid \mathbf{m}_{\mathcal{G}}^{\mathrm{pr}}(o)\big) = 0\,. \tag{21}$$

Note that we assume the set $\mathcal{M}_{\mathcal{G}}^{\mathrm{pr}}$ of price-relevant characteristics to be time-invariant. From a theoretical *a posteriori* viewpoint, this condition is less restrictive than it first appears since we only request that $\mathcal{M}_{\mathcal{G}}^{\mathrm{pr}}$ be finite. It may thus well assemble the whole set of characteristics which prove to be price-relevant in at least one of the time periods considered. If a characteristic is price-irrelevant in a certain period of time, the corresponding hedonic function will just neglect it.

The central aspect of this reformulation of the hedonic econometric model is the postulated time-dependency of the hedonic function $h_{\mathcal{G}}^t$. Fixed items are sold on the market for different prices at different points in time (this is what price statistics is all about), and the hedonic function mirrors these movements in how the market evaluates the inherent quality of an item. For an elementary aggregate where the hedonic econometric model holds, the quality-adjusted price evolution is thus fully represented by the evolution of the hedonic function over time. An appropriate comparison of the hedonic functions in a base and a current period may thus be seen as an implementation of an elementary price index measuring pure price change.

Before we proceed to the formulation of hedonic elementary population indices based on the idea just described, we propose simplifying the notation by 'randomising' the hedonic econometric model introduced above. Imagine a random draw from all the items of an elementary aggregate $\mathcal{G}^t$ at time t and denote by $\mathbf{M}^t$ the random vector of price-relevant characteristics and P^t the random variable representing the price of the drawn

item.[5] By the hedonic econometric model, the relationship between $\mathbf{M}^t$ and P^t is given by

$$P^t = h_{\mathcal{G}}^t(\mathbf{M}^t) + \epsilon^t \tag{22}$$

where the random error ϵ^t has $E\epsilon^t = 0$ for all $t \in \mathcal{T}$ and is assumed to be independent of $\mathbf{M}^t$. Within this additive error model, the hedonic function $h_{\mathcal{G}}^t$ therefore is exactly the conditional mean

$$h_{\mathcal{G}}^t(\mathbf{m}) = E(P^t \mid \mathbf{M}^t = \mathbf{m}), \tag{23}$$

and the conditional distribution $\mathbb{P}(P^t \mid \mathbf{M}^t)$ depends on $\mathbf{M}^t$ only through $h_{\mathcal{G}}^t$.

2.4.2 *Simple hedonic elementary population indices*

In Section 2.3 we identified two approaches for defining elementary population indices either as the ratio of some average prices (17) or as some average of the price ratios (18). Both of these population indices were defined on the restricted set $\mathcal{G}^0 \cap \mathcal{G}^1$ of items available in both the base and the current period. With this restriction, it was ensured that the compared prices belonged to identical goods and thus that the qualities of the items compared were equal. Consequently, the measured price evolution was not subject to any bias due to quality change.

Hedonic elementary price indices adhere to the paradigm of fixed reference qualities, but they do not rely on a fixed set of items for which prices need to be available in both time periods. Once the hedonic function for a certain time period is determined, it is able to deliver imputed prices for virtually any vector of price-relevant characteristics and as such for any item quality. The idea of hedonic elementary price indices is now to fix the reference quality of an elementary aggregate through vectors of price-relevant characteristics. With the help of the hedonic functions, the reference qualities are mapped to corresponding prices that can ultimately be compared using one of the two elementary price index approaches introduced above.

The simplest form of a hedonic elementary price index takes just one vector $\boldsymbol{\mu}^*$ of price-relevant characteristics as reference

5 The distribution of the random variable P^t corresponds to the distribution of prices denoted by $\tilde{p}^t(\mathcal{G})$ in Section 2.3, and the expectation $E(P^t)$ is one possible implementation of $\mu(\tilde{p}^t(\mathcal{G}))$.

quality. Irrelevant of the type of elementary population index used, this yields the index formula

$$HEPI^{0:1}(\mathcal{G}) = \frac{h^1_{\mathcal{G}}(\mu^*)}{h^0_{\mathcal{G}}(\mu^*)} = \frac{E(P^1 \mid \mathbf{M}^1 = \mu^*)}{E(P^0 \mid \mathbf{M}^0 = \mu^*)} \tag{24}$$

which we call *simple hedonic elementary population index*.[6] It relates the imputed price of the reference quality μ^* at time 1 to its imputed price at time 0.[7]

The open question here is how μ^* should be defined. The most obvious approach is to take some mean vector of price-relevant characteristics of the items available at the base or the current period. Formally, this gives us either $\mu^* = E\mathbf{M}^0$ or $\mu^* = E\mathbf{M}^1$ and with (24) corresponding implementations of the simple hedonic elementary population index.

The disadvantage of both of these implementations is that they asymmetrically favour the quality spectrum of the elementary aggregate at either the base or the current period. We therefore propose to work with a generalised reference quality distribution, represented by a random vector $\mathbf{M}$. The most natural choice for this reference distribution would probably be a mixture of $\mathbf{M}^0$ and $\mathbf{M}^1$, i.e.

$$\mathbb{P}_{\mathbf{M}} = g\,\mathbb{P}_{\mathbf{M}^0} + (1-g)\,\mathbb{P}_{\mathbf{M}^1}\,, \tag{25}$$

with $\mathbb{P}_{\mathbf{M}}$, $\mathbb{P}_{\mathbf{M}^0}$ and $\mathbb{P}_{\mathbf{M}^1}$ being the probability measures of $\mathbf{M}$, $\mathbf{M}^0$ and $\mathbf{M}^1$, respectively, and $g \in (0,1)$. If we set $\mu^* = E\mathbf{M}$, we get with $g = 0$ or $g = 1$ the two implementations of simple hedonic elementary population indices already introduced above and with $g = 1/2$ a sensible candidate of an index that symmetrically incorporates the quality spectrum in both the base and the current period.[8]

6 Hill and Melser (2008) use the term 'characteristics price index' for this type of index formula.

7 Technically, the index (24) is only well-defined if μ^* lies in $\mathfrak{m}^{pr}_{\mathcal{G}}(\mathcal{G}^0) \cap \mathfrak{m}^{pr}_{\mathcal{G}}(\mathcal{G}^1)$, i.e. in the domains of both $h^0_{\mathcal{G}}$ and $h^1_{\mathcal{G}}$. If this is not the case, a minimal requirement is that both hedonic functions can be extended to a domain including μ^*. This is normally not a problem in practice if a regression approach is chosen that allows for reasonable out-of-sample prediction.

8 For the special case of parametric hedonic functions, Brachinger (2002) introduces simple hedonic elementary population indices of the type (24) under the name of 'true hedonic price indices'. He distinguishes explicitly the implementations obtained when $\mu^* = E\mathbf{M}$ with $g = 0$, 1, or $1/2$. Referring to their orientation towards either the base, the current, or both periods simultaneously, he calls these implementations 'true hedonic Laspeyres price index', 'true hedonic Paasche price index', and 'true hedonic adjacent periods price index', respectively.

2.4.3 *Full hedonic elementary population indices*

Simple hedonic elementary population indices evaluate the 'distance' of the two hedonic functions in the base and in the current period at just one single quality point $\boldsymbol{\mu}^*$. Although this is certainly a valid practice, there are ways of better exploiting the full spectrum of the reference quality distribution and of obtaining a more representative index value.

One such way is to transform the whole reference quality distribution with the help of the two hedonic functions and to compare the resulting price distributions using the approaches described in Section 2.3.2. If we take the expectation as measure of location μ, the population indices (17) and (18) translate into *full hedonic elementary population indices* defined by

$$HEPI^{0:1}(\mathcal{G}) = \frac{\mathrm{E}h^1_{\mathcal{G}}(\boldsymbol{M})}{\mathrm{E}h^0_{\mathcal{G}}(\boldsymbol{M})} \tag{26}$$

and

$$HEPI^{0:1}(\mathcal{G}) = \mathrm{E}\left[\frac{h^1_{\mathcal{G}}(\boldsymbol{M})}{h^0_{\mathcal{G}}(\boldsymbol{M})}\right] . \tag{27}$$

In both cases, the expectations are built over the whole range of $\boldsymbol{M}$ and cover thus the reference quality distribution as a whole.[9]

We see that the distribution of $\boldsymbol{M}$ in principle does not need to be related to either $\boldsymbol{M}^0$ or $\boldsymbol{M}^1$, although a mixture like (25) is probably still the most reasonable choice. The minimum assumption to be made is that the range of $\boldsymbol{M}$ is contained in the domain of both $h^0_{\mathcal{G}}$ and $h^1_{\mathcal{G}}$. Note that, following (23), we have

$$\begin{aligned} \mathrm{E}h^t_{\mathcal{G}}(\boldsymbol{M}) &= \mathrm{E}_{\boldsymbol{M}}(\mathrm{E}_{P^t \mid \boldsymbol{M}^t}(P^t \mid \boldsymbol{M})) \\ &= \int_{\mathbb{R}^{K_{\mathcal{G}}}} \left[\int_{\mathbb{R}} p \, d\mathbb{P}_{P^t|\boldsymbol{M}^t}(p \mid \boldsymbol{m})\right] d\mathbb{P}_{\boldsymbol{M}}(\boldsymbol{m}) . \end{aligned} \tag{28}$$

for $t \in \{0, 1\}$. Here, $\mathbb{P}_{P^t|\boldsymbol{M}^t}$ stands for the probability measure of the conditional distribution of P^t given $\boldsymbol{M}^t$, and $\mathrm{E}_{P^t \mid \boldsymbol{M}^t}$ is the expectation with respect to this probability measure. Moreover, $\mathbb{P}_{\boldsymbol{M}}$ is the probability measure respective to the distribution

9 Diewert et al. (2009) showed that for the widely used special case of log-linear hedonic functions and under certain assumptions for the reference quality distribution used (which are satisfied when $g = 0$ or $g = 1$), both the simple and the full hedonic elementary indices are equivalent.

of $\mathbf{M}$, and $E_{\mathbf{M}}$ is its expectation. If one considers continuous random variables and vectors, equation (28) can be rewritten as

$$Eh_{\mathcal{G}}^{t}(\mathbf{M}) = \int_{\mathbb{R}^{K_{\mathcal{G}}}} \left[\int_{\mathbb{R}} p \, \frac{f_{(P^t,\mathbf{M}^t)}(p,\mathbf{m})}{f_{\mathbf{M}^t}(\mathbf{m})} \, dp \right] f_{\mathbf{M}}(\mathbf{m}) \, d\mathbf{m} \tag{29}$$

with $f_{(P^t,\mathbf{M}^t)}$ being the common probability density of P^t and $\mathbf{M}^t$, $f_{\mathbf{M}^t}$ the marginal density of $\mathbf{M}^t$ and, finally, $f_{\mathbf{M}}$ the density of $\mathbf{M}$. It can be seen that for this equation to be well-defined, the support of $f_{\mathbf{M}}$ needs to be contained in the support of $f_{\mathbf{M}^t}$ for $t \in \{0,1\}$. In other words, for each vector $\mathbf{m} \in \mathbb{R}^{K_{\mathcal{G}}}$ with $f_{\mathbf{M}^t}(\mathbf{m}) = 0$, it is necessary that $f_{\mathbf{M}}(\mathbf{m}) = 0$. This has to be taken into consideration when the reference quality $\mathbf{M}$ is chosen. In particular, $\mathbb{P}_{\mathbf{M}}$ must not attribute a positive probability to any set of characteristics vectors that does not have a positive probability with respect to $\mathbb{P}_{\mathbf{M}^0}$ and $\mathbb{P}_{\mathbf{M}^1}$ as well, i.e. within the populations available in both the base and current period.

In practice, therefore, it is even useful to assume that $\mathbb{P}_{\mathbf{M}^0}$ and $\mathbb{P}_{\mathbf{M}^1}$ attribute a positive probability to any non-discrete set of vectors in the characteristics space, i.e. the cartesian product of the ranges of all price-relevant characteristics. This ensures that out-of-sample-prediction is possible, and thus there are no formal restrictions on the distribution of reference characteristics $\mathbf{M}$.

2.4.4 *Universal formulae for hedonic elementary price indices*

In the last two sections, we introduced alternative definitions of a hedonic elementary population index. There the expectation operator was used as a special choice of a measure of location. There is, however, no a priori reason for this restriction. A natural generalisation of this approach results if we admit transformations of the price distributions. The expectation of the transformed price distribution characterises the location of this distribution. A measure of location of the original price distribution then results from backtransforming the expectation of the transformed price distribution.

Based on these reflections, the full hedonic elementary population indices (26) and (27) can be generalised to

$$HEPI^{0:1}(\mathcal{G}) = \frac{\varphi^{-1}(E\varphi(h_{\mathcal{G}}^{1}(\mathbf{M})))}{\varphi^{-1}(E\varphi(h_{\mathcal{G}}^{0}(\mathbf{M})))} \tag{30}$$

and

$$HEPI^{0:1}(\mathcal{G}) = \varphi^{-1}\left(\mathrm{E}\left[\varphi\left(\frac{h^1_{\mathcal{G}}(\mathbf{M})}{h^0_{\mathcal{G}}(\mathbf{M})}\right)\right]\right) \tag{31}$$

where φ is a continuous and injective function that maps a connected subset of $\mathbb{R}$ to $\mathbb{R}$ and φ^{-1} is its inverse.

With respect to the usual elementary price index formulae, three particular φ-functions play an important role. These are the identity, the hyperbolic transformation $\varphi(x) = x^{-1}$ as well as the natural logarithm $\varphi(x) = \ln x$. We will see below that depending on the choice of φ among these alternatives, the well-known hedonic elementary sample indices can be derived. Note that both definitions, (30) and (31), coincide if $\varphi(x) = \ln x$. This is due to the linearity of the expectation and the properties of the natural logarithm.[10]

We propose with (30) and (31) two universal prototypes of hedonic elementary population indices that leave, however, some degrees of freedom for the choice of φ and of the reference distribution of $\mathbf{M}$. We argued already that the latter is reasonably defined as a (symmetric) mixture of the base and the current period characteristics. However, there is no evident argumentation that favours either choice of φ except for the coincidence of both formulae if $\varphi(x) = \ln x$. Beer (2007a) discussed this question – a variant of what is called the *price index problem* in the literature (see e.g. Hill and Melser, 2008) – in the light of the well-known axiomatic approach to statistical price indices (Eichhorn and Voeller, 1976; Eichhorn, 1978) and managed to prove that (30) is preferable to (31) since it satisfies all proposed index axioms if $\varphi(\lambda x) = \varphi(\lambda) + \varphi(x)$ or $\varphi(\lambda x) = \varphi(\lambda)\varphi(x)$ for all $\lambda, x \in \mathbb{R}$. This latter condition, however, holds for all three φ-functions proposed above, so the axiomatic approach does not seem to be sufficient for choosing one 'best' universal hedonic

10 Silver and Heravi (2007) use exactly (30) with $\varphi(x) = \ln x$ as the definition of a 'Jevons' population index, although just for the case of conventional (i.e. non-hedonic) elementary price indices. In fact, we could well rewrite (17) and (18) in an analogous way as

$$EPI^{0:1}(\mathcal{G}) = \frac{\varphi^{-1}(\mathrm{E}\varphi(P^1))}{\varphi^{-1}(\mathrm{E}\varphi(P^0))}$$

and

$$EPI^{0:1}(\mathcal{G}) = \varphi^{-1}\left(\mathrm{E}\left[\varphi\left(P^1/P^0\right)\right]\right)$$

with P^0 and P^1 being the base and current period prices of the same item randomly drawn from the reference set $\mathcal{G}^0 \cap \mathcal{G}^1$.

elementary population index. Hence, there are obviously other arguments that need to be considered.

If there are no theoretical reasons that determine the choice of a specific population index, the introduction of the hedonic econometric model described by (20) offers at least the possibility to gauge the candidate index formulae according to the statistical properties of their estimators. Why not opt for the population index which can be estimated with highest statistical precision? Comparing the lengths of confidence intervals of estimators for the various population indices could therefore offer a new approach to address the price index problem. We are going to follow this idea through in the next two sections.

2.5 CONFIDENCE INTERVALS OF HEDONIC PRICE INDICES

2.5.1 *Hedonic imputation indices*

So far, we always remained on the abstract level of index definitions and population indices which are, as we repeatedly stressed, economic parameters that cannot directly be observed and eventually need to be estimated. For this purpose, we first turn towards the estimation of the time-varying hedonic functions using an appropriate regression approach. Any estimate $\hat{h}^t_{\mathcal{G}}$ of the hedonic function $h^t_{\mathcal{G}}$ can be seen as the result of a mapping

$$\begin{array}{rcccl} \mathfrak{h} & : & \mathbb{R}^{N^t}_{\geqslant 0} \times \mathbb{R}^{N^t \times K_{\mathcal{G}}} & \longrightarrow & \mathcal{H} \\ & & (\mathbf{P}^t, \mathbf{M}^t) & \longmapsto & \hat{h}^t_{\mathcal{G}} := \mathfrak{h}[\mathbf{P}^t, \mathbf{M}^t] \end{array} \tag{32}$$

where $\mathbf{P}^t = (P^t_1, \ldots, P^t_{N^t})$ denotes the random vector of sampled prices, $\mathbf{M}^t$ is the respective $N^t \times K_{\mathcal{G}}$ random matrix of the price-relevant characteristics $(\mathbf{M}^t_1, \ldots, \mathbf{M}^t_{N^t})$ observed in period t, and $\mathcal{H} := \{h : \mathbb{R}^{K_{\mathcal{G}}} \longrightarrow \mathbb{R}_{\geqslant 0}\}$ denotes the admissible hedonic functions applicable to the given elementary aggregate $\mathcal{G}$.[11]

Assume that $\hat{h}^t$ ($t \in \{0, 1\}$) are estimators of the hedonic functions h^t based on regression of item characteristics to prices in period 0 and 1, respectively. Then, relying on an i.i.d. sample of reference characteristics vectors $\mathbf{M}_1, \ldots, \mathbf{M}_N$ where $\mathbf{M}_n \overset{L}{\sim} \mathbf{M}$

11 As the elementary aggregate is supposed to be fixed, we are going to lighten the notation by dropping the index $\mathcal{G}$ from here on wherever possible.

for all $n \in \{1, \ldots, N\}$, sample versions of the universal population indices defined by (30) and (31) are given by

$$\widehat{HEPI}^{0:1} = \frac{\varphi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\varphi(\hat{h}^1(\mathbf{M}_n))\right)}{\varphi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\varphi(\hat{h}^0(\mathbf{M}_n))\right)} \tag{33}$$

and

$$\widehat{HEPI}^{0:1} = \varphi^{-1}\left(\frac{1}{N}\sum_{n=1}^{N}\varphi\left(\frac{\hat{h}^1(\mathbf{M}_n)}{\hat{h}^0(\mathbf{M}_n)}\right)\right), \tag{34}$$

respectively.

From these two formulae, by choosing φ among the alternatives mentioned above, we get the five hedonic elementary sample indices displayed in Table 2 which are hedonic counterparts to the elementary sample indices summarised in Table 1. We recognise that the elementary index formulae most widely used in practice (see e.g. ILO et al., 2004, paras. 20.38–45) prove to be estimators of the population indices (30) and (31). Among these are the indices attributed to Dutot, Jevons, and Carli, and the one that is called 'Harmonic Carli' here. Moreover, we find a 'Harmonic Dutot' sample index which to our knowledge does not appear in the literature. Note that when using $\varphi(x) = \ln x$ both general sample indices (33) and (34) lead to the Jevons elementary price index formula.

Once the φ-function is fixed and the distribution of $\mathbf{M}$ is defined, the single remaining influence factor that determines the performance of the hedonic elementary sample indices (33) and (34) and thus eventually the statistical quality of the index estimates is the regression approach used to estimate the hedonic functions. As we already discussed in Section 2.2, estimating the relationship between characteristics and price is a purely statistical issue with no *a priori* restriction on the functional form or regression approach to choose. As Triplett (2004, p. 186) stated, 'Any empirical form that fits the data is consistent with the theory.' So the entire repertoire of regression analysis can be applied to find an approach that best fits the data and delivers price predictions with the highest possible precision. The only point to remember is that estimated hedonic functions are normally used to perform some out-of-sample predictions where they should still provide plausible estimates.

In practice, the prevalent regression approaches for estimating hedonic functions are linear, semi-log and double-log models which perform well for many data sets. Curry et al. (2001)

Formula	Transformation	Sample index	Index type
(33)	$\varphi(x) = x$	$\widehat{HEPI}_{D}^{0:1} = \frac{\sum_{n=1}^{N} \hat{h}^1(\mathbf{M}_n)}{\sum_{n=1}^{N} \hat{h}^0(\mathbf{M}_n)}$	Dutot
	$\varphi(x) = \ln x$	$\widehat{HEPI}_{J}^{0:1} = \frac{\sqrt[N]{\prod_{n=1}^{N} \hat{h}^1(\mathbf{M}_n)}}{\sqrt[N]{\prod_{n=1}^{N} \hat{h}^0(\mathbf{M}_n)}}$	Jevons
	$\varphi(x) = x^{-1}$	$\widehat{HEPI}_{HD}^{0:1} = \frac{\left(\sum_{n=1}^{N} (\hat{h}^1(\mathbf{M}_n))^{-1}\right)^{-1}}{\left(\sum_{n=1}^{N} (\hat{h}^0(\mathbf{M}_n))^{-1}\right)^{-1}}$	'Harm. Dutot'
(34)	$\varphi(x) = x$	$\widehat{HEPI}_{C}^{0:1} = \frac{1}{N} \sum_{n=1}^{N} \frac{\hat{h}^1(\mathbf{M}_n)}{\hat{h}^0(\mathbf{M}_n)}$	Carli
	$\varphi(x) = \ln x$	$\widehat{HEPI}_{J}^{0:1} = \sqrt[N]{\prod_{n=1}^{N} \frac{\hat{h}^1(\mathbf{M}_n)}{\hat{h}^0(\mathbf{M}_n)}}$	Jevons
	$\varphi(x) = x^{-1}$	$\widehat{HEPI}_{HC}^{0:1} = \left(\frac{1}{N} \sum_{n=1}^{N} \left(\frac{\hat{h}^1(\mathbf{M}_n)}{\hat{h}^0(\mathbf{M}_n)}\right)^{-1}\right)^{-1}$	'Harm. Carli'

Table 2: Hedonic elementary sample indices.

were among the few authors who argued for a more flexible functional form, although the neural network approach they tested at the example of TVs did not show to be favourable to the linear or semi-log models.[12] Beer (2007a) investigated the use of conventional models compared to a partial least squares approach in an empirical study on used cars data. There, the winning model in terms of lowest bootstrap aggregate prediction error was an adaptive semi-log approach where individual regressions with automated variable selection and outlier detection were carried out and used for prediction for each of the car models in the sample. In the housing example to be presented in Section 2.6, we are going to rely on a semi-log approach with variable selection and outlier detection. Technically, each of these regression models is nothing else than a specific choice of $\mathfrak{h}$, transforming observations of prices and price-relevant characteristics in each period to estimates of the hedonic functions.

12 The failure of their neural network approach was particularly due to its instability on out-of-sample predictions.

2.5.2 *Bootstrapped confidence intervals for hedonic imputation indices*

It seems ambitious to analyse the statistical qualities of HEPI estimators given the potential complexity of the hedonic functions and the generally unknown form of the reference characteristics distribution. The most promising approach in this situation is certainly to address this question with an appropriate bootstrap procedure. Conditioned on the functional form of the hedonic regression, bootstrap confidence intervals deliver an insight on how precisely hedonic elementary sample indices estimate the corresponding population indices.

Drawing from Beer (2007b,a), we suggest to use a *wild bootstrap* approach as it was described by Davidson and Flachaire (2000) – at least in cases where a linear regression approach is chosen for estimating the hedonic function and thus standardised residuals are available. The main advantage of the wild bootstrap is its ability to cope with heteroscedastic error terms which are *a priori* not excluded in a setting defined by the hedonic econometric model (20) together with (21). With $\hat{I}^{0:1}$ being one of the five hedonic elementary price indices considered in Table 2, a confidence interval for the corresponding population index is obtained using the following resampling procedure:

1. For each time period $t \in \{0, 1\}$, estimate the hedonic regression function $\hat{h}^t := \mathfrak{h}[\mathbf{P}^t, \mathbf{M}^t]$ from the respective samples of price and characteristics observations in each period and compute the corresponding hedonic elementary price index $\widehat{I}^{0:1}$ using the sample of reference characteristics vectors $\mathbf{M}_1, \ldots, \mathbf{M}_N$.

2. For each bootstrap replication $s = 1, \ldots, S$,
   a) for each time period $t \in \{0, 1\}$,
      i. obtain simulated residuals $\epsilon^t_{\star sn} := r^t_n v_{\star sn}$ $(n = 1, \ldots, N^t)$ by multiplying each of the standardised residuals r^t_n of the hedonic regression models by an independent realisation $v_{\star sn}$ of a random variable that follows a Rademacher distribution;
      ii. compute simulated price values $p^t_{\star sn} := \hat{h}^t(\mathbf{M}^t_n) + \epsilon^t_{\star sn}$ use them to estimate the simulated hedonic function $\hat{h}^t_{\star s} := \mathfrak{h}[\mathbf{p}^t_{\star s}, \mathbf{M}^t]$, where

$$\mathbf{p}^t_{\star s} = (p^t_{\star s1}, \ldots, p^t_{\star sN^t}).$$

b) Calculate a simulated hedonic price index $\hat{I}^{0:1}_{\star s}$ using the simulated hedonic functions $\hat{h}^{t}_{\star s}$ and the same sample of reference characteristics vectors $\mathbf{M}_1, \ldots, \mathbf{M}_N$ as in Step 1. Compute the estimation error $\zeta^{0:1}_{\star s} := \hat{I}^{0:1}_{\star s} - \hat{I}^{0:1}$.

3. The increasingly ordered estimation errors $\zeta^{0:1}_{\star[s]}$, $s = 1, \ldots, S$, finally allow to compute the $(1 - 2\alpha)$ confidence interval as

$$\left[\widehat{I}^{0:1} - \zeta^{0:1}_{\star[(S+1)(1-\alpha)]}, \widehat{I}^{0:1} - \zeta^{0:1}_{\star[(S+1)\alpha]}\right].$$

Obviously, the number S of bootstrap replications is chosen such that $(S + 1)\alpha$ is an integer.

It should be noted here that by leaving the reference characteristics vectors $\mathbf{M}_1, \ldots, \mathbf{M}_N$ fixed, it is assumed that $\mathbb{P}_{\mathbf{M}}$ puts positive probability on this discrete, fixed and known set of reference characteristics vectors only. This might be somewhat too restrictive in theory and could easily be resolved by introducing resampled reference characteristics vectors in Step 2b of the algorithm. However, a previous study carried out by one of the authors (see Beer, 2007a, p. 136) showed almost no additional variance stemming from this source, so taking a fixed reference set should be sufficient in practice.

2.5.3 *The special case of time dummy hedonic indices*

We shall close this section by a comment on the *time dummy variable method*, which is a widely used alternative to the hedonic imputation indices discussed above (see e.g. Griliches, 1971, p. 59, Silver and Heravi, 2003, pp. 280–1, Triplett, 2004, p. 48–55, Diewert et al., 2009, or Hill, 2011). There, the price and characteristics data of both the base and the current period are pooled and the price-relevant characteristics $\mathbf{m} = (m_1, \ldots, m_K)'$ are supplemented by a time dummy variable t. Then a joint parametric hedonic function $h^{\{0,1\}}_{\mathcal{G}}$ is estimated on the basis of the pooled sample. From the estimated hedonic function $\hat{h}^{\{0,1\}}_{\mathcal{G}}$ two period-specific hedonic functions $\hat{h}^{t}_{\mathcal{G}}$ $(t = 0, 1)$ are easily recovered through

$$\hat{h}^{t}_{\mathcal{G}}(\mathbf{m}) := \hat{h}^{\{0,1\}}_{\mathcal{G}}(\mathbf{m}, t)\,. \tag{35}$$

These can be plugged into all hedonic elementary sample index formulae presented above.

An interesting situation emerges if we adopt the semi-log functional form for estimating the hedonic function. Then the relevant regression equation is given by

$$\ln P = \beta_0 + \delta\, t + \sum_{k=1}^{K} \beta_k M_k + \epsilon \tag{36}$$

and the estimated hedonic functions $\hat{h}_g^t$ can be written as

$$\hat{h}_g^t(\mathbf{m}) = \exp\left(\hat{\beta}_0 + \hat{\delta}\, t + \sum_{k=1}^{K} \hat{\beta}_k\, m_k\right). \tag{37}$$

with $\hat{\beta}_0, \ldots, \hat{\beta}_k$ and $\hat{\delta}$ being the OLS estimates of the corresponding coefficients in (36). Obviously,

$$\hat{h}_g^1(\mathbf{m}) = \exp\ \hat{\delta} \times \hat{h}_g^0(\mathbf{m}) \tag{38}$$

for all $\mathbf{m}$, and all of the sample index formulae listed in Table 2 reduce to $\widehat{HEPI}^{0:1} = \exp\ \hat{\delta}$. They are thus completely independent of the reference quality distribution used.

Although the property of independence just described sounds appealing, we agree with, e.g., Diewert et al. (2009) who argue in favour of hedonic imputation indices. In contrast to the time dummy approach, they have the advantage of not imposing any constraint on the functional form and eventually on the parameters of the hedonic functions. We are convinced that the flexibility of the functional form is important and therefore that any technical restrictions should be avoided.

2.6 HEDONIC INDICES FOR SINGLE-FAMILY DWELLINGS

2.6.1 *The data*

The data used for the present analysis were kindly provided by Wüest & Partner, an international consultancy firm for real estate. Transaction prices of single-family dwellings and their corresponding characteristics were collected for the Swiss canton of Zurich from banks, insurances, and other real estate agencies. The collected data are organized in 44 subsequent quarterly data sets, spanning from the first quarter of 2001 to the fourth quarter of 2011. The number of sampled observations per quarter ranges from 137 in the first quarter 2002 to 411 in the fourth quarter 2010, covering in average more than 50% of the transactions occurred in the relevant area.

The set $\mathcal{M}_{\mathcal{G}}^{pr}$ of price-relevant characteristics is defined by means of the following observed characteristics: age (*age*; in years), volume (*vol*; in cubic meters), surface of the land surrounding the property (*land*; in square meters), status (*status*; low–medium or *sup*erior)[13], condition (*cond*; poor–reasonable or *exc*ellent), micro location of the house within the municipality (*micro*; bad–medium or *good*), house type (*type*; semi-detached or *det*ached), number of rooms (*rooms*), the macro location of the house within the canton (*macro*; centre, *s*outh, or *n*orth), and number of parking spaces (*park*). Each quarterly data set is considered as containing price and characteristics information for items sampled from the induced elementary aggregate $\mathcal{G}^t = \mathcal{G}^t(\mathcal{M}_{\mathcal{G}})$, where $\mathcal{M}_{\mathcal{G}}$ are the distinguishing characteristics of 'single-family dwellings', $\mathcal{M}_{\mathcal{G}}^{pr} \subset \mathcal{M}_{\mathcal{G}}$ and $t = 0, \ldots, 44$.

All the computations carried out in the present paper were done using the free software environment R (Version 2.15.1, Windows, 64-bit; see R Core Team, 2012). In particular, the hedonic elementary price indices were computed by means of the HEPI package (see Beer, 2007a) available at R-Forge[14].

2.6.2 *Specification and estimation of quarterly hedonic functions*

Based on the considered set of price-relevant characteristics, a semi-log hedonic function was independently estimated for each quarter ($t = 1, \ldots, 44$) following the model equation

$$\begin{aligned}\log(p_i^t) = {} & \beta_0 + \beta_1 \log(age_i^t) + \beta_2 \log(vol_i^t) \\ & + \beta_3 \log(land_i^t) + \beta_4 status\_sup_i^t + \beta_5 cond\_exc_i^t \\ & + \beta_6 micro\_good_i^t + \beta_7 type\_det_i^t + \beta_8 rooms_i^t \\ & + \beta_9 macro\_s_i^t + \beta_9 macro\_n_i^t + \beta_{10} park_i^t \\ & + \beta_{11} \log(land_i^t) type\_det_i^t \\ & + \beta_{12} \log(land_i^t) micro\_good_i^t \\ & + \beta_{13} \log(land_i^t) macro\_s_i^t \\ & + \beta_{14} \log(land_i^t) macro\_n_i^t + \epsilon_i^t, \end{aligned} \tag{39}$$

where ϵ_i^t represents an error term satisfying the hypothesis stated in (21). The choice of this functional form rests on the following considerations: As already mentioned in Section 2.5.1,

13 All categorical variables were dummy-coded with the first mentioned category being the null case.

14 `http://r-forge.r-project.org/projects/hepi/`

Figure 1: Quarterly hedonic elementary price indices from the first quarter 2001 to the fourth quarter 2011. The five considered hedonic indices plotted together (top left) and individually with their corresponding bootstrapped 95% confidence intervals (remaining plots).

taking the logarithm of the transaction price is a well-established approach in the hedonic literature, mainly intended to obtain a better fit and mitigate potential heteroscedasticity problems. In contrast to other hedonic house price models such as those adopted by Goodman and Thibodeau (1995), Fletcher et al. (2000), and Stevenson (2004b), we decided to include the age variable in logarithmic form as well. This choice is primarily motivated by the inclusion of the status and condition dummies, which adequately account for both the age-related vintage and renovation effects described by Goodman and Thibodeau (1995). Due to a high variation in the land price across different geographical locations and house types, three interaction terms were additionally considered to take into account these effects. Note that, referring to the principles outlined in section Section 2.2.2, the chosen specification of the hedonic function does not rest on any economic theory but tries to accommodate best the considered data.

After a first estimation of the 44 regression models, an average of 7% of the observations were detected as unduly influential based on the DFFITS criterion and removed for the fol-

lowing analyses. Koenker's studentised version of the Breush-Pagan test against heteroscedasticity was subsequently applied to each regression model, revealing that 33 out of 44 linear models were plagued by heteroscedasticity at the 5% level. This result suggested that the use of a price-logarithmic hedonic function to address the heteroscedasticity problem was, in general, ineffective.

Since the observed heteroscedasticity could, however, also be due to a model misspecification, a heteroscedasticity-robust Ramsey Regression Equation Specification Error Test (RESET) was carried out to exclude this possibility. For each quarter the second, third, and fourth power of the fitted values obtained by means of equation (39) were included in the extended regression. The joint significance of the added regressors was then tested using a Wald test relying on the HC covariance matrix estimator recommended by Long and Ervin (2000). For each quarter the null hypothesis of a correct model specification could not be rejected at the 5% level.

The presence of heteroscedasticity having been confirmed, a heteroscedasticity-robust Wald test (again based on the covariance matrix estimator recommended by Long and Ervin (2000)) was finally applied to eliminate variables that were both individually and jointly not significant at the standard level. We refrained from using even more sophisticated variable selection methods for fear of overfitting the data, knowing that the hedonic functions were going to be used to some out-of-sample prediction. The adjusted R^2 statistics of the retained models range from 0.57 to 0.84.

2.6.3 *Computation of hedonic imputation indices and bootstrapped confidence intervals*

The first quarter of 2001 was chosen as base period $t = 0$ for all indices computed. A different reference sample $\mathbf{M}_1, \ldots, \mathbf{M}_{N^0+N^t}$ was compiled for each quarter by pooling the base period's observed characteristics with the observed characteristics in the relevant current period. The hedonic functions estimated in Section 2.6.2 were then used to compute the five hedonic elementary sample indices presented in Table 2 for each period. The wild bootstrap approach described in Section 2.5.2 was carried out with $S = 999$ replications resulting in a 95% confidence

Dutot	Harm. Dutot	Carli	Harm. Carli	Jevons
1.1591	1.0027	1.0234	0.9942	1.0000

Table 3: Average confidence interval lengths as percentage of the Jevons interval length.

interval for each hedonic index.[15] The index values and confidence intervals obtained are shown in Fig. 1.

The behaviour of the five indices is similar with all of them showing a 40 to 50 percent price increase for the eleven years under investigation. Interestingly, although all confidence intervals appear to be comparable, only the interval lengths of the Harmonic Carli, Harmonic Dutot, and Jevons sample indices are roughly the same on average while the interval lengths of the Carli and Dutot sample indices are substantially larger (see Table 3). Conditioned on the semi-log functional form used for estimating the hedonic functions, the former three sample indices seem to show the lowest variance due to sampling from the price and characteristics distributions in the base and reference period. Particularly noteworthy is the poor performance of the Dutot price index. Combining these empirical results with the axiomatic reflections cited in Section 2.4.4, it is the Harmonic Dutot and the Jevons index formulae that seem to be preferable when choosing the hedonic elementary population index to be used. This is consistent with the results obtained by Beer (2007a) for his study on used cars, where, additionally, the Jevons index was seen to be least sensitive to the (mis-)specification of the hedonic functional form.

Taking note of these observations, it seems to us thus that the Jevons hedonic elementary price index is the most attractive universal formula from both a theoretical and empirical point of view. It is clear, however, that for certain data sets, the empirical behaviour of the index estimators may be different from the results obtained here. We appreciate thus any further research on this topic.

15 Note that the simulated price values in Step 2(a)ii of the bootstrapping algorithm were calculated as $\log(p^t)_{\star sn} := \hat{h}^t(\boldsymbol{M}_n^t) + \epsilon_{\star sn}^t$ in order to accommodate the price-logarithmic functional form (see also Beer, 2007b, p. 89).

2.7 SUMMARY

We started the present piece of work with a formal definition of elementary aggregates. From the outset, much emphasis was put on the duality between elementary aggregates and their distinguishing characteristics. The latter played a central role when we translated the hedonic hypothesis known from the literature into a hedonic econometric model.

After discussing the two fundamental concepts of elementary price indices, we defined a list of hedonic elementary population indices reaching from simple indices where the entire quality range of an object was represented by a single vector of price-relevant characteristics to two universal index formulae showing much flexibility in how the price distributions in the base and current period are compared. As population indices are unobservable economic parameters, there is a need for sample indices acting as appropriate estimators of these parameters. We were able to show that most of the index formulae used in practice could be naturally derived within the proposed theoretical framework. The established framework additionally allowed to implement a procedure for computing bootstrapped confidence intervals for hedonic elementary indices.

Neither of the universal formulae of hedonic elementary population indices proposed in this paper is completely determined. So the user needs to make some further decisions in order to obtain a concrete target population index and eventually a corresponding sample index formula for practical applications. To this end, bootstrapped confidence interval lengths were used to compare the considered hedonic elementary indices.

Based on axiomatic reflections and empirical computations, we argued that the most attractive candidate of a hedonic elementary price index for an elementary aggregate $\mathcal{G}$ and for any given reference quality distribution $\mathbb{P}_{\mathbf{M}}$ was

$$HEPI^{0:1}(\mathcal{G}) = \exp\left(\mathrm{E}\left[\ln\left(\frac{h^1_{\mathcal{G}}(\mathbf{M})}{h^0_{\mathcal{G}}(\mathbf{M})}\right)\right]\right). \tag{40}$$

An estimator of this index is given by the Jevons formula

$$\widehat{HEPI}^{0:1}_{J} = \sqrt[N]{\prod_{n=1}^{N}\frac{\hat{h}^1(\mathbf{M}_n)}{\hat{h}^0(\mathbf{M}_n)}} \tag{41}$$

based on estimates $\hat{h}^0$ and $\hat{h}^t$ of the hedonic functions and with $\mathbf{M}_1, \ldots, \mathbf{M}_N$ sampled symmetrically from the price-relevant characteristics in both the base and the current period.

ACKNOWLEDGEMENTS

Parts of this study were developed within the project 'Specification, axiomatic foundation, and estimation of hedonic price indices' funded by the Swiss National Science Foundation as well as the Swiss Federal Statistical Office. The authors acknowledge these fundings. Moreover, we express our gratitude to Wüest & Partner for providing the data for the empirical study.

3

ASYMPTOTIC PROPERTIES OF IMPUTED HEDONIC PRICE INDICES IN THE CASE OF LINEAR HEDONIC FUNCTIONS

3.1 INTRODUCTION

To measure the price changes of goods and services relative to a base period, price statisticians have defined different sorts of price indices. To cope with a possible quality variation of the goods across different time periods, the hedonic approach has established itself, in the last ten years, as the most appropriate method for computing quality-adjusted price indices. The basis of the hedonic approach is the hedonic hypothesis: Each good is considered as a bundle of characteristics, and its price solely depends on these characteristics. Unfortunately, economic theory provides no guiding theory on the choice of the hedonic index, thus making the investigation of the statistical properties of these indices of primary importance.

A great amount of research has been carried out to appropriately model the relationship between the price of a good and its characteristics. To a lesser extent, researchers have focused on hedonic price indices, solely suggesting alternative formulae to their computation. These alternative formulae use, in general, the hedonic hypothesis to impute the price of goods in classical price index formulae[1]. Surprisingly, the obtained formulae are mainly used as descriptive statistical measures, thus completely neglecting their probabilistic nature. In a recent paper, Brachinger et al. (2012) used a bootstrap approach to evaluate the statistical properties of different elementary hedonic price indices, empirically showing, in particular, that some indices seem to have smaller confidence interval lengths than others.

The aim of the present paper is to use a standard probabilistic approach to determine the asymptotic properties of single imputed, double imputed, and characteristic hedonic price indices. This approach should provide a better understanding of the theoretical parameter that a hedonic index tries to estimate.

The present paper is structured as follows. Section 3.2 reviews

1 Hedonic price indices based on time dummy variables are not analyzed in the present paper since they do not rely on traditional price index formulae.

the most common imputed hedonic price indices. The convergence in probability of hedonic price indices is then analyzed in Section 3.3. Section 3.4 concludes the paper.

3.2 HEDONIC IMPUTED PRICE INDICES: A REVIEW

Three main approaches are considered in the present paper to compute quality-adjusted price indices: single imputed, double imputed, and characteristics methods. As mentioned in the footnote, time dummy price indices are not analyzed. Moreover, only the hedonic counterpart of the classical Laspeyres, Paasche, and Fisher price indices are considered. The adopted terminology and the following definitions are based on Hill (2011).

Let $\mathbf{P}^t := (P_1^t, ..., P_{n_t}^t)' \in \mathbb{R}^{n_t}$ and $\mathbf{X}^t := (\mathbf{x}_1^t, ..., \mathbf{x}_{n_t}^t)' \in \mathbb{R}^{n_t \times K}$ denote a vector of random prices and a matrix of random characteristics in period t, respectively. The hedonic hypothesis states that for each time period, the price of a good solely depends on its characteristics. The corresponding statistical model is given by

$$P_i^t = f^t(\mathbf{x}_i^t) + \epsilon_i^t = f^t(x_{i1}^t, ..., x_{iK}^t) + \epsilon_i^t, \ i = 1, ..., n_t,$$

where x_{ij}^t is the j-th characteristic of good i in period t, and the function f^t describes how the characteristics interact to build the price. The function f^t is usually called the hedonic regression function, or simply the hedonic function. We denote the hedonic function estimated in period t by $\hat{f}^t$. The set of observed characteristics is constant through time, i.e., is the same for each time period $t = 1, ..., T$. The number of goods observed in period t is denoted by n_t, and it is assumed that $n_t \geqslant K$. The variable ϵ_i^t represents a stochastic error term.

One major consequence due to the above statistical model is that, even if we identify a good with its characteristics, the price randomly varies. This means that for exactly the same good, we can observe different prices. This implies, in turn, that the quantity variable used to compute classical price indices is not appropriate for the hedonic approach. It is therefore recommended to consider each good as being unique and set the quantity variable q_i^t equal to 1 for each good appearing in the traditional price index formulae.

3.2.1 *Single imputed hedonic price indices*

Single imputed hedonic price indices use the statistical model implied by the hedonic hypothesis to impute the prices of each good considered in one time period according to the hedonic function estimated in the other time period. Estimated prices in one period are then compared to imputed prices in the other period. If we define the quality of a good by its vector of characteristics, it becomes evident that the quality of the goods does not change between periods.
Let $\widehat{HIL}^{si}_{0,t}$, $\widehat{HIP}^{si}_{0,t}$, and $\widehat{HIF}^{si}_{0,t}$ denote the estimators of the Laspeyres, Paasche, and Fisher single imputed hedonic price indices, respectively. They are defined as

$$\widehat{HIL}^{si}_{0,t} = \frac{\sum_{i=1}^{n_0} \hat{f}^t(\mathbf{x}_i^0)}{\sum_{i=1}^{n_0} P_i^0}$$
$$\widehat{HIP}^{si}_{0,t} = \frac{\sum_{i=1}^{n_t} P_i^t}{\sum_{i=1}^{n_t} \hat{f}^0(\mathbf{x}_i^t)}$$
$$\widehat{HIF}^{si}_{0,t} = \sqrt{(\widehat{HIL}^{si}_{0,t})^{1/2}(\widehat{HIP}^{si}_{0,t})^{1/2}},$$

where 0 and t represent the base and current time periods, respectively. The base period is chosen among the time periods $t = 1, ..., T$. These estimators are random variables that attempt to estimate the unknown population hedonic price index.

3.2.2 *Double imputed hedonic price indices*

In contrast to single imputed indices, double imputed hedonic price indices impute prices for both time periods. Once the hedonic functions for the two periods have been estimated, the set of characteristics in one period is evaluated according to the hedonic function estimated in the other period. By construction, this guarantees that the quality of the goods does not change between periods, and so prices can be directly compared.
Let $\widehat{HIL}^{di}_{0,t}$, $\widehat{HIP}^{di}_{0,t}$ and $\widehat{HIF}^{di}_{0,t}$ denote the estimators of the Las-

peyres, Paasche, and Fisher double imputed hedonic price indices, respectively. They are defined as

$$\widehat{HIL}^{di}_{0,t} = \frac{\sum_{i=1}^{n_0} \hat{f}^t(\mathbf{x}_i^0)}{\sum_{i=1}^{n_0} \hat{f}^0(\mathbf{x}_i^0)}$$
$$\widehat{HIP}^{di}_{0,t} = \frac{\sum_{i=1}^{n_t} \hat{f}^t(\mathbf{x}_i^t)}{\sum_{i=1}^{n_t} \hat{f}^0(\mathbf{x}_i^t)}$$
$$\widehat{HIF}^{di}_{0,t} = \sqrt{(\widehat{HIL}^{di}_{0,t})^{1/2}(\widehat{HIP}^{di}_{0,t})^{1/2}},$$

where 0 and t represent the base and current time periods, respectively. The base period is chosen among the time periods $t = 1, ..., T$.

3.2.3 *Characteristic hedonic price indices*

Instead of imputing prices for each good in a given period, characteristic hedonic price indices compute a representative good for one time period, and then impute the price of this characteristic good using the estimated hedonic functions in the two periods. The characteristic good is thought to appropriately represent the set of goods in one time period and is usually defined as being the mean vector of the characteristics. Also, in this case, since only the characteristic good is considered, quality does not change across periods, and prices are directly comparable. Let $\widehat{HIL}^{ch}_{0,t}$, $\widehat{HIP}^{ch}_{0,t}$, and $\widehat{HIF}^{ch}_{0,t}$ denote the estimators of the Laspeyres, Paasche, and Fisher characteristic hedonic price indices, respectively. They are defined as

$$\widehat{HIL}^{ch}_{0,t} = \frac{\hat{f}^t(\bar{\mathbf{x}}^0)}{\hat{f}^0(\bar{\mathbf{x}}^0)}$$
$$\widehat{HIP}^{ch}_{0,t} = \frac{\hat{f}^t(\bar{\mathbf{x}}^t)}{\hat{f}^0(\bar{\mathbf{x}}^t)}$$
$$\widehat{HIF}^{ch}_{0,t} = \sqrt{(\widehat{HIL}^{ch}_{0,t})^{1/2}(\widehat{HIP}^{ch}_{0,t})^{1/2}},$$

where $\bar{\mathbf{x}}^t := (\bar{x}^t_1, ..., \bar{x}^t_K) = (\frac{1}{n_t}\sum_{i=1}^{n_t} x^t_{i1}, ..., \frac{1}{n_t}\sum_{i=1}^{n_t} x^t_{iK})$ represents the mean vector of the characteristics.

3.3 CONVERGENCE IN PROBABILITY

Although widely employed, the above defined indices have been used as empirical quantities, without considering their statistical properties. To derive such properties, the hedonic function

f_t considered in the hedonic hypothesis must be specified and an estimation technique accordingly adopted. In the present paper, the following linear hedonic function

$$f_t(x_{i1}^t, ..., x_{iK}^t) := (\mathbf{x}_i^t)'\boldsymbol{\beta}^t = \\ = \beta_0^t + \beta_1^t x_{i1}^t + ... + \beta_K^t x_{iK}^t, \; i = 1, ..., n_t \tag{42}$$

is assumed in each time period.

The following proposition identifies the theoretic indices toward which the above defined hedonic price indices converge.

Theorem 3.3.1. *Let* $(P_i^t, \boldsymbol{x}_i^t)$, $i = 1, ..., n_t$ *be a random sample of* n_t *independent random variables belonging to period* t $(t = 1, ..., T)$. *We assume that in each time period, the characteristics' vector* $\boldsymbol{x}_i^t$ *follow the same probability distribution with finite mean:* $\boldsymbol{x}_i^t \sim \boldsymbol{x}^t$ *and* $\boldsymbol{\mu}_{\boldsymbol{x}^t} := \mathbb{E}(\boldsymbol{x}^t) < +\infty \; \forall i$. *If the usual hypotheses* [2] *hold in each time period for the linear hedonic model in (42), then*

i) $\widehat{HIL}_{0,t}^{si}$, $\widehat{HIL}_{0,t}^{di}$, *and* $\widehat{HIL}_{0,t}^{ch}$ *converge in probability toward* $\frac{\boldsymbol{\mu}'_{x^0}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{x^0}\boldsymbol{\beta}^0}$.

ii) $\widehat{HIP}_{0,t}^{si}$, $\widehat{HIP}_{0,t}^{di}$, *and* $\widehat{HIP}_{0,t}^{ch}$ *converge in probability toward* $\frac{\boldsymbol{\mu}'_{x^t}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{x^t}\boldsymbol{\beta}^0}$.

Proof. Some of the following equalities are explained in 3.5 at the end of the document.
i) The convergence in probability of $\widehat{HIL}_{0,t}^{si}$ is first established:

$$\begin{aligned}
\operatorname*{plim}_{n_0, n_t \to +\infty} \widehat{HIL}_{0,t}^{si} &= \operatorname*{plim}_{n_0, n_t \to +\infty} \frac{\sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t}{\sum_{i=1}^{n_0} P_i^0} = \\
&= \operatorname*{plim}_{n_0 \to +\infty} \frac{\sum_{i=1}^{n_0} (\mathbf{x}_i^0)'(\operatorname{plim}_{n_t \to \infty} \hat{\boldsymbol{\beta}}^t)}{\sum_{i=1}^{n_0} P_i^0} = \\
&= \operatorname*{plim}_{n_0 \to +\infty} \frac{\sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\boldsymbol{\beta}^t}{\sum_{i=1}^{n_0} P_i^0} = \\
&= \frac{\operatorname{plim}_{n_0 \to +\infty} \frac{1}{n_0} \sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\boldsymbol{\beta}^t}{\operatorname{plim}_{n_0 \to +\infty} \frac{1}{n_0} \sum_{i=1}^{n_0} P_i^0} = \\
&= \frac{\mathbb{E}((\mathbf{x}_i^0)'\boldsymbol{\beta}^t)}{\mathbb{E}(P_i^0)} = \frac{\mathbb{E}((\mathbf{x}_i^0))'\boldsymbol{\beta}^t}{\mathbb{E}_{\mathbf{x}_i^0}(\mathbb{E}(P_i^0|\mathbf{x}_i^0))} = \\
&= \frac{\boldsymbol{\mu}'_{x^0}\boldsymbol{\beta}^t}{\mathbb{E}_{\mathbf{x}_i^0}((\mathbf{x}_i^0)'\boldsymbol{\beta}^0)} = \frac{\boldsymbol{\mu}'_{x^0}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{x^0}\boldsymbol{\beta}^0}.
\end{aligned}$$

2 See Greene (2011), page 92.

For double imputed indices, we have

$$\begin{aligned}
\operatorname*{plim}_{n_0,n_t\to+\infty} \widehat{\mathrm{HIL}}^{\mathrm{di}}_{0,t} &= \operatorname*{plim}_{n_0,n_t\to+\infty} \frac{\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t}{\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^0} = \\
&= \operatorname*{plim}_{n_0\to+\infty} \frac{\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'(\operatorname{plim}_{n_t\to\infty}\hat{\boldsymbol{\beta}}^t)}{\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^0} = \\
&= \operatorname*{plim}_{n_0\to+\infty} \frac{\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'\boldsymbol{\beta}^t}{\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^0} = \\
&= \frac{\operatorname{plim}_{n_0\to+\infty}\frac{1}{n_0}\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'\boldsymbol{\beta}^t}{\operatorname{plim}_{n_0\to+\infty}\frac{1}{n_0}\sum_{i=1}^{n_0}(\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^0} = \\
&= \frac{\mathbb{E}((\mathbf{x}_i^0)'\boldsymbol{\beta}^t)}{\mathbb{E}((\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^0)} = \frac{\mathbb{E}((\mathbf{x}_i^0))'\boldsymbol{\beta}^t}{\mathbb{E}_{\mathbf{X}^0}(\mathbb{E}((\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^0|\mathbf{X}^0))} = \\
&= \frac{\boldsymbol{\mu}'_{\mathbf{x}^0}\boldsymbol{\beta}^t}{\mathbb{E}_{\mathbf{x}_i^0}((\mathbf{x}_i^0)'\boldsymbol{\beta}^0)} = \frac{\boldsymbol{\mu}'_{\mathbf{x}^0}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{\mathbf{x}^0}\boldsymbol{\beta}^0}.
\end{aligned}$$

For $\widehat{\mathrm{HIL}}^{\mathrm{ch}}_{0,t}$, we simply have

$$\operatorname*{plim}_{n_0,n_t\to+\infty} \widehat{\mathrm{HIL}}^{\mathrm{ch}}_{0,t} = \frac{(\operatorname{plim}_{n_0\to+\infty}\bar{\mathbf{x}}^0)'(\operatorname{plim}_{n_t\to+\infty}\hat{\boldsymbol{\beta}}^t)}{(\operatorname{plim}_{n_0\to+\infty}\bar{\mathbf{x}}^0)'(\operatorname{plim}_{n_0\to+\infty}\hat{\boldsymbol{\beta}}^0)} = \frac{\boldsymbol{\mu}'_{\mathbf{x}^0}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{\mathbf{x}^0}\boldsymbol{\beta}^0}.$$

ii) The demonstrations for the Paasche indices are similar:

$$\begin{aligned}
\operatorname*{plim}_{n_0,n_t\to+\infty} \widehat{\mathrm{HIP}}^{\mathrm{si}}_{0,t} &= \operatorname*{plim}_{n_0,n_t\to+\infty} \frac{\sum_{i=1}^{n_t} P_i^t}{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^0} = \\
&= \operatorname*{plim}_{n_t\to+\infty} \frac{\sum_{i=1}^{n_t} P_i^t}{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'(\operatorname{plim}_{n_0\to\infty}\hat{\boldsymbol{\beta}}^0)} = \\
&= \operatorname*{plim}_{n_t\to+\infty} \frac{\sum_{i=1}^{n_t} P_i^t}{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\boldsymbol{\beta}^0} = \\
&= \frac{\operatorname{plim}_{n_t\to+\infty}\frac{1}{n_t}\sum_{i=1}^{n_t} P_i^t}{\operatorname{plim}_{n_t\to+\infty}\frac{1}{n_t}\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\boldsymbol{\beta}^0} = \\
&= \frac{\mathbb{E}(P_i^t)}{\mathbb{E}((\mathbf{x}_i^t)'\boldsymbol{\beta}^0)} = \frac{\mathbb{E}_{\mathbf{x}_i^t}(\mathbb{E}(P_i^t|\mathbf{x}_i^t))}{\mathbb{E}((\mathbf{x}_i^t))'\boldsymbol{\beta}^0} = \\
&= \frac{\mathbb{E}_{\mathbf{x}_i^t}((\mathbf{x}_i^t)'\boldsymbol{\beta}^t)}{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^0} = \frac{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^0}.
\end{aligned}$$

and

$$\begin{aligned}
\operatorname*{plim}_{n_0,n_t\to+\infty} \widehat{\mathrm{HIP}}^{\mathrm{di}}_{0,t} &= \operatorname*{plim}_{n_0,n_t\to+\infty} \frac{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^t}{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^0} = \\
&= \operatorname*{plim}_{n_t\to+\infty} \frac{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^t}{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'(\operatorname{plim}_{n_0\to\infty}\hat{\boldsymbol{\beta}}^0)} = \\
&= \operatorname*{plim}_{n_t\to+\infty} \frac{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^t}{\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\boldsymbol{\beta}^0} = \\
&= \frac{\operatorname{plim}_{n_t\to+\infty}\frac{1}{n_t}\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^t}{\operatorname{plim}_{n_t\to+\infty}\frac{1}{n_t}\sum_{i=1}^{n_t}(\mathbf{x}_i^t)'\boldsymbol{\beta}^0} = \\
&= \frac{\mathbb{E}((\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^t)}{\mathbb{E}((\mathbf{x}_i^t)'\boldsymbol{\beta}^0)} = \frac{\mathbb{E}_{\mathbf{X}^t}(\mathbb{E}((\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^t|\mathbf{X}^t))}{\mathbb{E}((\mathbf{x}_i^t))'\boldsymbol{\beta}^0} = \\
&= \frac{\mathbb{E}_{\mathbf{x}_i^t}((\mathbf{x}_i^t)'\boldsymbol{\beta}^t)}{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^0} = \frac{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^0}.
\end{aligned}$$

and

$$\operatorname*{plim}_{n_0,n_t\to+\infty} \widehat{\mathrm{HIP}}^{\mathrm{ch}}_{0,t} = \frac{(\operatorname{plim}_{n_t\to+\infty}\overline{\mathbf{x}}^t)'(\operatorname{plim}_{n_t\to+\infty}\hat{\boldsymbol{\beta}}^t)}{(\operatorname{plim}_{n_t\to+\infty}\overline{\mathbf{x}}^t)'(\operatorname{plim}_{n_0\to+\infty}\hat{\boldsymbol{\beta}}^0)} = \frac{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^0}.$$

□

The convergence in probability of the Laspeyres and Paasche hedonic price indices can then be used to establish the convergence in probability of the Fisher index.

Corollary 1. *The Fisher hedonic price indices* $\widehat{\mathrm{HIF}}^{\mathrm{si}}_{0,t}$, $\widehat{\mathrm{HIF}}^{\mathrm{di}}_{0,t}$, *and* $\widehat{\mathrm{HIF}}^{\mathrm{ch}}_{0,t}$ *converge in probability toward*

$$\sqrt{\left(\frac{\boldsymbol{\mu}'_{x^0}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{x^0}\boldsymbol{\beta}^0}\right)^{1/2}\left(\frac{\boldsymbol{\mu}'_{x^t}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{x^t}\boldsymbol{\beta}^0}\right)^{1/2}}.$$

Proof. For single imputed hedonic price indices, we have

$$\begin{aligned}
\operatorname*{plim}_{n_0,n_t\to+\infty} \widehat{\mathrm{HIF}}^{\mathrm{si}}_{0,t} &= \sqrt{(\operatorname*{plim}_{n_0,n_t\to+\infty}\widehat{\mathrm{HIL}}^{\mathrm{si}}_{0,t})^{1/2}(\operatorname*{plim}_{n_0,n_t\to+\infty}\widehat{\mathrm{HIP}}^{\mathrm{si}}_{0,t})^{1/2}} = \\
&= \sqrt{\left(\frac{\boldsymbol{\mu}'_{\mathbf{x}^0}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{\mathbf{x}^0}\boldsymbol{\beta}^0}\right)^{1/2}\left(\frac{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^t}{\boldsymbol{\mu}'_{\mathbf{x}^t}\boldsymbol{\beta}^0}\right)^{1/2}}.
\end{aligned}$$

For double imputed and characteristic hedonic price indices, the proof is identical. □

The following corollary is directly implied by Proposition 3.3.1 Corollary 1:

Corollary 2. *Laspeyres, Paasche, and Fisher price indices are asymptotically equivalent in the case of single imputed, double imputed, and characteristics hedonic price indices.*

We have proved that hedonic price indices converge in probability toward a non-linear function of the statistical models' parameters. Interestingly, each estimator of the individual parameters possess a normal asymptotic distribution (see Greene (2011) for details):

- $\hat{\boldsymbol{\beta}}^0 \overset{a}{\sim} N(\boldsymbol{\beta}^0, \frac{(\sigma^0)^2}{n_0} Q_{X_0}^{-1}),\ Q_{X_0}^{-1} := \text{plim}_{n_0 \to +\infty} \frac{1}{n_0}(X_0' X_0)^{-1}.$
- $\hat{\boldsymbol{\beta}}^t \overset{a}{\sim} N(\boldsymbol{\beta}^t, \frac{(\sigma^t)^2}{n_t} Q_{X_t}^{-1}),\ Q_{X_t}^{-1} := \text{plim}_{n_t \to +\infty} \frac{1}{n_t}(X_t' X_t)^{-1}.$
- $\hat{\mu}_{x_j^0} \overset{a}{\sim} N(\mu_{x_j^0}, \sigma^2_{x_j^0}/n_0),\ j = 1, ..., K.$
- $\hat{\mu}_{x_j^t} \overset{a}{\sim} N(\mu_{x_j^t}, \sigma^2_{x_j^t}/n_t),\ j = 1, ..., K.$

Unfortunately, even for these well-known distributions, it seems infeasible to derive the asymptotic distribution of the indices, even (unrealistically) assuming stochastic independence among time periods and/or regressors.

3.4 CONCLUSIONS

The asymptotically equivalence of single imputed, double imputed, and characteristics hedonic price indices has been established in the case of goods possessing a linear hedonic function. This result appears to be quite important, since it implies that the price index problem tends to vanish in large samples, thus alleviating an uncomfortable situation price statisticians have to face.
Despite their importance, the obtained results must also be carefully placed in the context in which estimation of hedonic models takes place. In the case of a non-linear hedonic regression function, in fact, our results are generally not valid, even for continuous hedonic functions. An important case is represented by log-linear hedonic models, which are commonly used to model the hedonic prices of housing goods. When inverse transformation is applied to obtain the imputed prices in the

original scale, the mean of a non-linear function has to be calculated to obtain the theoretical price index, thus invalidating the results found.

An indirectly interesting result concerns the asymptotic distribution of the hedonic price indices. It seems unrealistic to analytically compute the asymptotic distribution of such indices, therefore suggesting the use of resampling methods to determine their distribution.

In the present paper, only a specific class of imputed hedonic indices has been analyzed. It could be interesting to conduct further research to determine whether other hedonic price indices converge in probability toward the same theoretical parameter and establish if their asymptotic distribution could be explicitly determined.

3.5 APPENDIX: ASYMPTOTIC PROPERTIES OF LINEAR HEDONIC MODELS

The following properties hold under the classical linear model hypothesis and the hypothesis assumed in Proposition 3.3.1. The employed terminology is borrowed from DasGupta (2011).

Property 1. *In any time period* t *and base period* 0*, we have*

$$\operatorname*{plim}_{n_t \to +\infty} \frac{\sum_{i=1}^{n_0} (x_i^0)' \hat{\beta}^t}{\sum_{i=1}^{n_0} P_i^0} = \frac{\sum_{i=1}^{n_0} (x_i^0)' \beta^t}{\sum_{i=1}^{n_0} P_i^0}.$$

Proof. We consider first the convergence in probability of a single term $(\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t$ as $n_t \to +\infty$. The probability distribution of the K-dimensional random variable $(\mathbf{x}_i^0)$ does not depend on n_t. It can therefore be considered as converging in probability toward itself as $n_t \to +\infty$: $\operatorname{plim}_{n_t \to +\infty}(\mathbf{x}_i^0) = (\mathbf{x}_i^0)$. Under the classical hypothesis of the linear regression model estimated in period t, we have that $\operatorname{plim}_{n_t \to +\infty} \hat{\boldsymbol{\beta}}^t = \boldsymbol{\beta}^t$. The multi-dimensional convergence preservation implies that $\operatorname{plim}_{n_t \to +\infty}(\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t = (\mathbf{x}_i^0)'\boldsymbol{\beta}^t$. Using again the convergence preservation, we obtain

$$\operatorname*{plim}_{n_t \to +\infty} \sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t = \sum_{i=1}^{n_0} \operatorname{plim}_{n_t \to +\infty} \left((\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t \right) = \sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\boldsymbol{\beta}^t.$$

Since the denominator $\sum_{i=1}^{n_0} P_i^0$ does not depend on n_t, we also have that $\operatorname{plim}_{n_t \to +\infty} \sum_{i=1}^{n_0} P_i^0 = \sum_{i=1}^{n_0} P_i^0$. Thus implying

$$\operatorname*{plim}_{n_t \to +\infty} \frac{\sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t}{\sum_{i=1}^{n_0} P_i^0} = \frac{\operatorname{plim}_{n_t \to +\infty} \sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\hat{\boldsymbol{\beta}}^t}{\operatorname{plim}_{n_t \to +\infty} \sum_{i=1}^{n_0} P_i^0} = \frac{\sum_{i=1}^{n_0} (\mathbf{x}_i^0)'\boldsymbol{\beta}^t}{\sum_{i=1}^{n_0} P_i^0}.$$

□

Remark 1. *Property 1 is also valid if the denominator is replaced with* $\sum_{i=1}^{n_0} (x_i^0)'\hat{\boldsymbol{\beta}}^0$*, since it represents a random variable not depending on* n_t*.*

Proof. If in time period t a linear hedonic function is assumed, then

i) $\mathbb{E}(P_i^t) = \boldsymbol{\mu}'_{\mathbf{x}^t} \boldsymbol{\beta}^t$

ii) $\mathbb{E}((\mathbf{x}_i^t)'\hat{\boldsymbol{\beta}}^t) = \boldsymbol{\mu}'_{\mathbf{x}^t} \boldsymbol{\beta}^t$.

□

Proof. It is assumed that $P_i^t = (x_i^t)'\beta^t + \epsilon_i^t$. According to the linear model hypotheses, we have that $\mathbb{E}(\epsilon_i^t | x_i^t) = 0$. A stronger form of exogeneity is not necessary, since the characteristics vectors x_i^t, $i = 1, ..., n_t$ are assumed to be independent in a given time period.
i) Using the law of iterated expectations, we have

$$\begin{aligned}\mathbb{E}(P_i^t) &= \mathbb{E}_{x_i^t}(\mathbb{E}(P_i^t | x_i^t)) = \mathbb{E}_{x_i^t}\left(\mathbb{E}\left((x_i^t)'\beta^t + \epsilon_i^t | x_i^t\right)\right) = \\ &= \mathbb{E}_{x_i^t}((x_i^t)'\beta^t) = \mu'_{x^t}\beta^t.\end{aligned}$$

ii) The second equality is slightly more complicated to demonstrate. Substituting the P^t with the statistical model in the classical OLS estimation formula $\hat{\beta}^t = ((X^t)'(X^t))^{-1}(X^t)'P^t$, we have

$$\hat{\beta}^t = \beta^t + ((X^t)'(X^t))^{-1}(X^t)'\epsilon_t,$$

with $\epsilon^t := (\epsilon_1^t, ..., \epsilon_{n_t}^t)'$. Using the exogeneity hypothesis, we have $\mathbb{E}(\epsilon^t | X^t) = 0$. We must condition on the whole characteristics matrix X^t to use the law of iterated expectations.

$$\begin{aligned}\mathbb{E}((x_i^t)'\hat{\beta}^t) &= \mathbb{E}_{X^t}(\mathbb{E}((x_i^t)'\hat{\beta}^t)|X^t) = \\ &= \mathbb{E}_{X^t}\left(\mathbb{E}\left((x_i^t)'(\beta^t + ((X^t)'(X^t))^{-1}(X^t)'\epsilon_t)|X^t\right)\right) = \\ &= \mathbb{E}_{X^t}\left(\mathbb{E}\left((x_i^t)'\beta^t + (x_i^t)'((X^t)'(X^t))^{-1}(X^t)'\epsilon_t|X^t\right)\right) = \\ &= \mathbb{E}_{x_i^t}((x_i^t)'\beta^t) = \\ &= \mu'_{x^t}\beta^t.\end{aligned}$$

□

4

A NEW APPROACH TO VARIABLE SELECTION IN THE PRESENCE OF MULTICOLLINEARITY: A SIMULATED STUDY WITH HEDONIC HOUSING DATA

4.1 INTRODUCTION

The hedonic approach considers each good as a bundle of characteristics. This bundle of characteristics is related to the price of the good through the so-called hedonic price function. Much research has been devoted to identifying the characteristics that determine the price-building process, or, alternatively, to establish if a specific set of characteristics significantly affects the price of a given good.

Two major problems arise when trying to identify price relevant characteristics in hedonic price regressions. First, no economic model is available to guide researches in selecting the characteristics. Second, multicollinearity is often detected among the characteristics, thus negatively influencing standard selection techniques. To partially overcome these difficulties, automated variable selection methods seem to provide a valid approach.

Automated variable selection methods allow, with a computer algorithm, the best subset of characteristics to be selected according to a specific selection criterion. This criterion is often represented by an objective function that has to be minimized. In early versions, automated techniques used the classical t or F tests as the selection criterion and suffered, therefore, from problems derived from carrying out several non-independent tests. Modern implementations, however, are based on selection criteria stemming from information theory, providing a better theoretical framework in which the selection takes place.

A major critique concerning automated selection methods is that a single 'best' model is selected, completely ignoring the other models. This criticism is particularly pertinent when several alternative models display values of the objective function similar to the objective function value of the selected model, a situation that typically occurs in the case of multicollinearity. A new approach proposed by Burnham and Anderson (2004) consists of computing the weight of each model with an infor-

mation criterion, and subsequently deducing the importance of each independent variable. The statistical analysis thus relies on the whole set of models, and this critique does not apply.
The paper is structured as follows. Section 4.2 introduces the automated variable selection techniques and selection criteria. The multimodel approach and the characteristics' importance are discussed in section 4.3. A new selection procedure based on the characteristics' importance is illustrated in section 4.4. The statistical measure used to gauge the variable selection techniques is then illustrated in section 4.5. Sections 4.6 and 4.7 define the simulation set-up and describe the data used, respectively. The results are shown in section 4.8. Section 4.9 concludes the paper.

4.2 STEPWISE SELECTION AND INFORMATION LOSS CRITERIA

Stepwise selection is a greedy algorithm that locally optimizes the criterion by successively nesting regression models. The present article considers only backward stepwise selection. Backward stepwise selection starts from a full model, where all the available variables are considered, and then drops the non-relevant variables. At each iteration of the algorithm, a single variable is identified by the selection criterion and then eliminated. The algorithm stops when dropping a variable does not significantly improve the value of the objective function.
Let $\mathcal{M} := \{M_1, ..., M_m\}$ be the set of candidate models and K the total number of characteristics. Since the focus is on backward stepwise selection, the models contained in $\mathcal{M}$ are all nested models of a full model M_F including all K characteristics. $M_i := \{y, x_1, ..., x_{K_i}\} \in \mathcal{M}$ denotes a regression model with dependent variable y and independent variables $x_1, ..., x_{K_i}$. It is assumed that the models are all applied to the same data set containing n observations. $I_L : \mathcal{M} \to \mathbb{R}$ denotes the objective function that has to be minimized. When a new step in the iteration process worsens the objective function's value more than a given tolerance level, the algorithm stops. The algorithm in 4.10 describes the different steps undertaken by backward stepwise selection to obtain a model that locally minimizes the loss function I_L. Since backward stepwise selection is a greedy algorithm, the search over the model set is not complete, i.e., the algorithm does not consider all possible 2^K models, therefore providing a possible globally sub-optimal solution.

Two different information loss functions I_L are considered in the present paper. The first criterion is the famous Akaike Information Criterion (AIC) proposed by Akaike (1973), which is defined as

$$\begin{aligned} AIC: \mathcal{M} &\longrightarrow \mathbb{R} \\ M_i &\longmapsto -2\log(L(\hat{\theta}|x_1, ..., x_{K_i}) + 2K_i, \end{aligned}$$

where L denotes the likelihood of the model. The second criterion is called the Bayesian Information Criterion (BIC) and was presented by Schwarz (1978). It is given by

$$\begin{aligned} BIC: \mathcal{M} &\longrightarrow \mathbb{R} \\ M_i &\longmapsto -2\log(L(\hat{\theta}|x_1, ..., x_{K_i}) + K_i \log(n). \end{aligned}$$

Both criteria are based on a log-likelihood function evaluated at the estimated parameters' value plus a term penalizing the model's complexity. In the case of a linear regression model with a normal distributed error term,
$-2\log(L(\hat{\theta}|y, x_1, ..., x_K)$ is simply equal to $n\log(\hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the maximum-likelihood estimated model variance.

4.3 THE MULTIMODEL APPROACH AND VARIABLES' IMPORTANCE

We consider as before a set $\mathcal{M} = \{M_1, ..., M_m\}$ of candidate models. For each model, we compute the values $IC_1, ..., IC_m$, where IC_i denotes either the AIC or BIC value of the i-th model. Let $IC_{Min} = \min_{i=1,...,m}\{IC(M_i)\}$ be the information criterion value of the best model. According to the multimodel approach proposed by Burnham and Anderson (2004), we define the models' weights as

$$\begin{aligned} w_{IC}: \mathcal{M} &\longrightarrow [0, 1] \\ M_i &\longmapsto \frac{\exp(-\frac{1}{2}(IC_i - IC_{Min}))}{\sum_{r=1}^{m}\exp(-\frac{1}{2}(IC_r - IC_{Min}))}. \end{aligned}$$

We clearly have that $0 \leqslant w_{IC}(M_i) \leqslant 1$ and $\sum_{i=1}^{m} w_{IC}(M_i) = 1$. From a Bayesian point of view, these weights represent the a posteriori probabilities of each model. As explained by Burnham and Anderson (2004), the defined AIC-based model weights put the following a priori probabilities on the set of candidate models:

$$P(M_i) = \frac{\exp\left(\frac{1}{2}K_i\log(n) - K_i\right)}{\sum_{i=1}^{m}\exp\left(\frac{1}{2}K_i\log(n) - K_i\right)}, \; i = 1, ..., m,$$

whereas BIC-based model weights implicitly assume that $P(M_i) = 1/m,\ i = 1, ..., m$.[1] One question that naturally arises when using the multimodel approach is how to determine the initial set $\mathcal{M}$ of alternative models. When no theory is available to specify a set of candidate models, a solution is to consider all possible models for the available variables. All-subset regression seems therefore indicated in this situation. All-subset regression is a branch and bound algorithm that, given an initial set of regressors, estimates all possible regression models. Assuming that the full model contains K regressors, the total number of models is equal to $2^K - 1$, where the empty model is a priori discarded.
To determine the importance of a given variable, Burnham and Anderson (2004) suggest the computation of the following variables' weights:

$$\begin{aligned} w_{IC}^{x,post} : \{x_1, ..., x_K\} &\longrightarrow [0, 1] \\ x_i &\longmapsto \sum_{j=1}^{m} w_{IC}(M_j) \mathbb{1}_{\{x_i \in M_j\}}, \end{aligned}$$

where $\mathbb{1}_{\{x_i \in M_j\}}$ is an indicator function. Even if a variable is not included in the best model according to the information criterion, the variable could nevertheless possess a high weight. If all models not including a given variable have weights near zero, then the weight of this variable is near one. Conversely, if the weights of the models not containing a given variable sum up to almost one, then the weight of this variable is near zero.

4.4 VARIABLES' WEIGHTS AND PRICE-RELEVANT CHARACTERISTICS

Since the main goal is to identify a single model that carefully describes the relation between the price of a good and its characteristics, we suggest using variables' weights to define an automated selection procedure; a given variable is included in the final model only if the variable reaches a minimum level of importance, i.e., the variable's weight is greater than a given value. The major difficulty is determining this minimum level. Two remarks are necessary in order to specify the minimum importance level. First, the importance level should not be based on the computed variables' weights, because such a selection

1 The interpretation of these a priori probabilities differs due to the statistical context in which the AIC and BIC are derived.

rule would tend to highlight data-specific features, thus causing poor model generalization. Second, the importance level should depend on the set of candidate models: If a variable is under-/over-represented in the set of models, a lower/higher importance level should be considered. According to these remarks, we assess the minimum level of importance with the following definition.

Definition. *For a given information criterion* IC *and a given set of candidate models* $\mathcal{M} = \{M_1, ..., M_m\}$*, the* ***a priori weight*** $w_{IC}^{x_i,ap}$ *of a variable* x_i*,* $i = 1, ..., K$ *is defined as*

$$w_{IC}^{x,ap} : \{x_1, ..., x_K\} \longrightarrow [0, 1]$$
$$x_i \longmapsto \sum_{j=1}^{m} P(M_j) \mathbb{1}_{\{x_i \in M_j\}},$$

where $P(M_i)$ *denotes the a priori probability of model* M_i*.*

The following corollary can be used to approximate the BIC a priori weights and save computing time:

Corollary. *If* $\mathcal{M}$ *is given by all possible regressors subsets and the number of regressors is large, then* $w_{BIC}^{x,ap}(x_i) \cong 0.5 \; \forall i$*.*

Proof. In the all-subset case, each variable is contained exactly in 2^{K-1} models. The a priori BIC-based variables weights are therefore equal to

$$w_{BIC}^{x,ap}(x_i) = \sum_{j=1}^{m} \frac{1}{m} \mathbb{1}_{\{x_i \in M_j\}} = \frac{2^{K-1}}{2^K - 1} \cong 0.5 \; \forall i \text{ when } K \gg 1.$$

□

Based on the a priori weights, the following variable selection rule has been implemented:

Selection Rule. *For a given information criterion IC and a given set of candidate models* $\mathcal{M}$*, a variable* $x_i \in M_F$ *is included in the final model if and only if* $w_{IC}^{x,post}(x_i) \geqslant w_{IC}^{x,ap}(x_i)$*.*

If the a posteriori weight of a given variable is greater than its a priori value, then the variable is included in the selected model. According to the previously stated selection rule, a variable is included in the selected model only if the BIC-based weight w_{BIC}^{x} is greater than 0.5.
The proposed selection method is easily implemented and, in contrast to AIC- and BIC-based stepwise regression, simultaneously considers the importance of a given variable in all models belonging to $\mathcal{M}$.

4.5 MEAN BALANCED ACCURACY

To assess the performance of the selection method, the aim of the regression must be clearly defined. Since the goal is to identify the independent variables that affect the dependent variable, the selection method has to be gauged according to its ability to identify the data-generating process or, at least, the model that most carefully approximates the data-generating process for a given class of models.
Let M_{sel} be a regression model selected with the preceding selection techniques. The task of identifying the original data-generating process can be viewed as a binary classification problem. Let IC be the number of informative variables correctly identified, IW the number of informative variables not included in the selected model, UC the number of uninformative variables correctly identified, and UW the number of uninformative variables included in the selected model. These four quantities represent the so-called confusion matrix. Using the confusion matrix, several accuracy measures are available. Recently, Brodersen et al. (2010) introduced the concept of balanced accuracy to measure the performance of a classification algorithm.[2] It is defined as

$$BA(M_{sel}|y, X) = \frac{1}{2}\left(\frac{IC}{IC + IW} + \frac{UC}{UC + UW}\right),$$

where y and X denote the observed dependent and independent variables, respectively. If S data sets $(y_1, X_1), ..., (y_S, X_S)$ from the same data-generating process are available, the Mean Balanced Accuracy (MBA) can be computed as follows

$$MBA = \frac{1}{S}\sum_{s=1}^{S} BA(M_{sel}|y_s, X_s).$$

4.6 SIMULATION STUDY

In real-world applications, the data-generating process is unknown. To gauge the performance of a given selection method, a data-generating process is thus simulated and the performance of the selection technique evaluated. We propose a simulation in the context of hedonic regression. First, a hedonic price function including a set of a priori chosen characteristics is esti-

2 Their paper empirically shows, in particular, the advantages of using balanced accuracy when unbalanced classification problems are considered.

mated. Second, the estimated hedonic function is used to simulate prices according to the characteristics. In the third step, an increasing number of noise variables inducing or not inducing multicollinearity is added to the set of characteristics. Finally, a selection algorithm is applied to a linear regression model based on the data set containing the simulated prices, the characteristics used to simulate the prices (informative variables), and the noise variables (uninformative variables).
Two remarks are of primary importance at this point. First, the 'true' data-generating process is contained in the initial full model. Second, we do not a priori specify the multicollinearity degree induced by noise variables. In the case of noise variables inducing multicollinearity, adding a new noise variable may not worsen the multicollinearity problem as much as the previously added noise variable.

4.6.1 *Price-generating process*

Let $(p_i, x_{i1}, ..., x_{ig})$, $g < K$ be the data set containing the price p_i of the i-th good in a given time period and the price-relevant characteristics $(x_{i1}, ..., x_{ig})$. The following process is used to simulate log-prices:

i) A log-linear hedonic model is assumed:

$$\log(p_i) = \beta_0 + \beta_1 x_{i1} + ... + \beta_g x_{ig} + \epsilon_i,$$

where ϵ_i represents a stochastic error term. Let $\widehat{\log(p_i)}$ denote the estimated log-price for the i-th observation.

ii) A simulated error term is added to the estimated log-prices to obtain the simulated log-prices according to the characteristics:

$$\log(p_i^*) = \widehat{\log(p_i)} + \epsilon_i^*,$$

where the simulated stochastic error term ϵ_i^* is randomly generated with a $N(0, \sigma^2)$ probability distribution.

The main problem is to adequately choose the variance σ^2 of the simulated error term. As suggested by Chong and Jun (2005), the error term's variance can be chosen to achieve a specific R^2 in the price-generating process. Let r^2 represent the R^2 value required from the log-price generating process. Computing the simulated error variance as

$$\sigma^2 = \left(\frac{1 - r^2}{r^2}\right) Var(\widehat{\log(\mathbf{p})})$$

guarantees that the R^2 of the data-generating process is equal to r^2. The term

$$Var(\widehat{\log(\mathbf{p})}) := Var((\widehat{\log(p_1)}, ..., \widehat{\log(p_n)}))$$

represents the empirical variance of the estimated log-prices.
To analyze how the selection method performs for different R^2 levels, two R^2 values were used to compute the simulated error variance: 50% and 90%. These levels were chosen since hedonic regressions rarely display R^2 values under 50%. For each of the R^2 levels, 1'000 data sets containing n observations were simulated.

4.6.2 *Noise variables and hedonic regression equations*

Once the log-prices have been simulated, a set of uninformative variables $(x_{i(g+1)}, ..., x_{iK})$ is added to the set of informative variables $(x_{i1}, ..., x_{ig})$. A selection method is then applied to the following full regression model

$$\log(p_i^*) = \beta_0 + \beta_1 x_{i1} + ... + \beta_g x_{ig} + \beta_{g+1} x_{i(g+1)} + ... + \beta_K x_{iK} + u_i,$$

where u_i is the full model stochastic error term.
To assess the performance of the selection method for a varying number of noise variables, the number of noise variables included in the model is progressively increased to obtain three different ratios of noise variables to the total number of variables considered. The ratios are 33%, 42%, and 50%. The noise variables added to the model have been chosen to induce or not induce multicollinearity in the full model.

4.7 THE DATA

The data used for the analysis were kindly provided by Wüest & Partner, an international real estate consultancy firm. Transaction prices of single-family dwellings and their corresponding characteristics were collected for the Swiss canton of Zurich from banks, insurances companies, and other real estate agencies. In the present study, the 320 transactions observed for the third quarter 2011 are considered.
The following characteristics were used to simulate the log-prices: age (*age*; in years), volume (*vol*; in cubic meters), surface of the land surrounding the property (*land*; in square meters), and the macro location of the house within the canton

Multicollinearity not induced				Multicollinearity induced			
age	1.2	1.2	1.3	*age*	42.5	43.2	43.2
vol	1.7	2.0	2.1	*vol*	1.6	6.7	10
land	1.6	1.6	1.7	*land*	1.6	1.8	15.6
$macro_s$	1.6	1.6	1.6	$macro_s$	1.6	1.6	1.6
$macro_n$	1.6	1.6	1.6	$macro_n$	1.6	1.6	1.6
$status_s$	1.2	1.2	1.2	age^2	239	241.4	243.1
$cond_e$	1.3	1.3	1.3	age^3	99.3	100	101.9
rooms		1.3	1.4	vol^2		7.2	9.15
$type_d$			1.2	$land^2$			13.5

Table 4: Variance inflation factors of the regression models. All categorical variables appear as factors in the regression equation with the first category defined as the reference group.

(*macro*; center, *s*outh, or *n*orth). In addition to these a priori chosen price-relevant characteristics, eight different noise variables were used. The noise variables not inducing multicollinearity are given by status (*status*; low–medium or *s*uperior), condition (*cond*; poor–reasonable or *e*xcellent), number of rooms (*rooms*), and the house type (*type*; semi-detached or *d*etached). The noise variable inducing multicollinearity are powers of the informative variables used to generate prices: age^2, age^3, vol^2, and $land^2$. Added noise variables that do not induce multicollinearity have also been chosen because they are potentially related to the informative variables. Clearly, this relation is non-linear: *cond* and *status* are typically age-related characteristics, whereas *rooms* and *type* are related to *vol* and *land*, respectively. This approach was chosen, since it allows a fairer comparison between noise variables not inducing and inducing multicollinearity.
The noise variables are progressively added to the log-price generating model to vary the percentage of noise variables to the total number of variables. Table 4 shows the regression models and the corresponding Variance Inflation Factor (VIF). As can be seen, when noise variables inducing multicollinearity are added to the data-generating process, the VIF values of the variables are greater than the usual recommended value of 5.

4.8 RESULTS

In this section, the results for the simulation set-up and the selection methods are described. All computations have been performed with the statistical software R (see R Core Team (2012) for further information).

4.8.1 *Backward stepwise selection and MBA*

The left side of Figure 2 shows the MBA values according to the number of noise variables not inducing and inducing multicollinearity and the R^2 levels, when a backward stepwise selection approach is used. Interestingly, independently of the information criterion, the performance of the backward stepwise selection method in the case of noise variables not inducing multicollinearity seems to be mostly unaffected by the number of added noise variables and the fit of the original data-generating process. As expected, the performance of the backward stepwise selection method is improved when the BIC criterion is used: Since the price-generating process is contained in the initial full model and is given by only a few big effects, the BIC criterion will asymptotically select the original data-generating process with a probability of 1 (see Burnham and Anderson (2004)). Nevertheless, both information criteria generally perform well when noise variables not inducing multicollinearity are used.
Not surprisingly, independently of the selection criterion, the selection method performs systematically worse when noise variables inducing multicollinearity are used. The MBA of the two stepwise selection methods seems to be equally affected by the introduction of noise variables inducing multicollinearity. Interestingly, adding a new noise variable inducing multicollinearity does not necessarily worsen the MBA value. This is probably because the multicollinearity induced by age^2 and age^3 is higher than for the vol^2 and $land^2$ variables (see Table 4), thus making it easier for the stepwise selection algorithm to identify these last two variables as added noise.

4.8.2 *Multimodel selection and MBA*

The findings resulting from the multimodel selection approach are illustrated in the right side of Figure 2. Analogously to the backward stepwise selection method, the performance of the

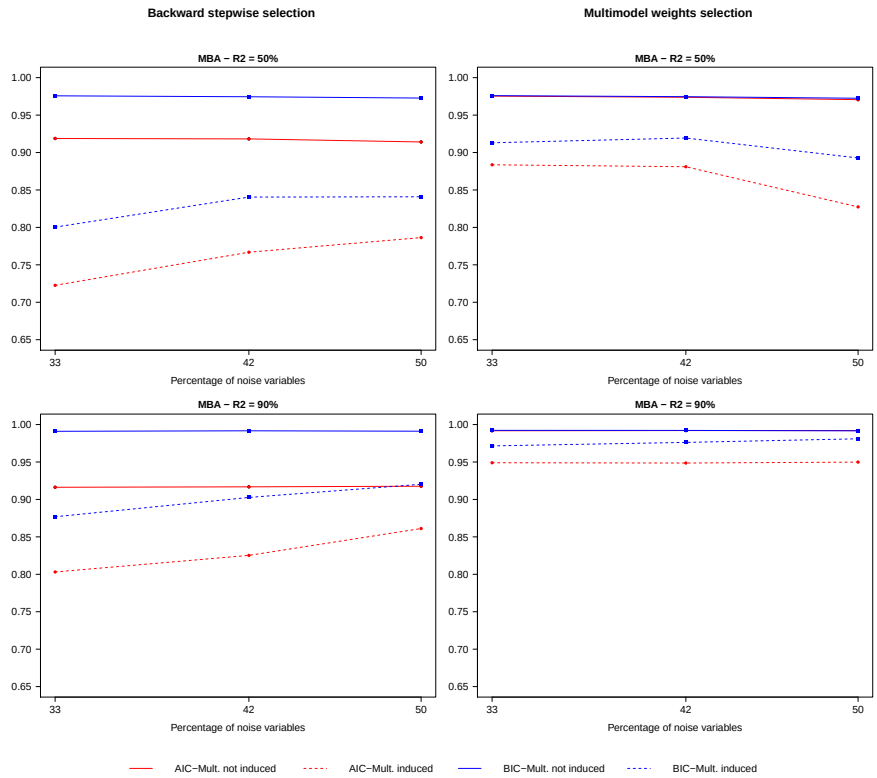

Figure 2: Mean balanced accuracy of backward stepwise selection and multimodel weights selection according to the number of noise variables and R^2 values.

multimodel selection approach is analyzed regarding the number of added noise variables and the two R^2 values.
When noise variables not inducing multicollinearity are added to the original data-generating process, the MBA, as for the previous selection method, seems to be mostly unaffected by the number of added noise variables and the fit of the data-generating process, and this occurs independently of the information criterion. This invariance, however, does not hold in the case of multicollinearity-inducing noise variables, with the MBA clearly higher for $R^2 = 90\%$. Moreover, when $R^2 = 50\%$, the number of added noise variables has an impact on the MBA values when the AIC is used, this time in the expected direction. Also remarkable is the performance of the AIC and BIC multimodel selection approaches for noise variables not inducing multicollinearity. In contrast to the previous selection methods, these approaches have virtually the same MBA values.
Very interestingly, the MBA for the multimodel approach is systematically better than that of the backward stepwise approach when multicollinearity is present, especially for a high R^2. This result stems from the very nature of the multimodel approach;

computing the variables' importance according to a well-calibrated set of candidate models allows the selection algorithm to gauge the relevance of each variable in each model in turn.

4.9 CONCLUSIONS

The present paper exclusively focused on identifying the data-generating process, assuming that this process is a nested version of a more general model in which all the available characteristics are included in the analysis. The findings in section 4.8 suggest that when multicollinearity is not present backward stepwise regression is a reliable method for identifying the original data-generating process, independently of its R^2 value. The performance of this selection method, however, is decreased in the case of noise variables inducing multicollinearity, thus requiring the use of an alternative selection approach. To solve this problem, a selection method based on a multimodel approach has been suggested. The proposed method clearly shows better performance in the case of multicollinearity and, surprisingly, seems to perform equally well with the AIC and BIC criteria. These two features are particularly important, since even if multicollinearity is present and the non-appropriate information criterion for the set-up is used, good MBA values are obtained.

Another advantage presented by the proposed multimodel method is that, due to the high MBA values, it seems reasonable to assume that the variance introduced by the selection method does not significantly increase the coefficients' variance. The usual formulas for the coefficients' standard deviations could therefore be used as approximations of the theoretical ones, thus making it possible to use classical test statistics once the regression model has been selected.

When the purpose of the hedonic regression is prediction, however, the relevance of the multimodel selection approach is questionable. A comparison of the out-of-sample prediction performance of the multimodel selected model with the prediction performance of an averaged model could be a subject for further research.

4.10 APPENDIX: BACKWARD SELECTION PSEUDO-ALGORITHM

Algorithm 1 Backward selection pseudo algorithm

Require: M_F, I_L, tol

$M \leftarrow M_F$, $g_M \leftarrow I_L(M_F)$, $\mathrm{iter} \leftarrow \mathrm{TRUE}$
while iter **do**
  for $i = 1$ **to** $|\{x_i \in M\}|$ **do**
    $M_i \leftarrow M \backslash \{x_i\}$, $g_i \leftarrow I_L(M_i)$
  end for
  $M_{drop} \leftarrow M_{i^*}$ such that $I_L(M_{i^*}) < I_L(M_i)\ \forall i$
  $g_{M_{drop}} \leftarrow I_L(M_{drop})$
  if $g_{M_{drop}} \geqslant g_M + \mathrm{tol}$ **then**
    $\mathrm{iter} \leftarrow \mathrm{FALSE}$
  else
    $M \leftarrow M_{drop}$, $g_M \leftarrow I_L(M_{drop})$
  end if
end while
return M

5

SELECTION OF REGRESSION METHODS IN HEDONIC PRICE MODELS BASED ON PREDICTION LOSS FUNCTIONS

5.1 INTRODUCTION

Research activity in hedonic price models can roughly be divided into two categories: Determining the impact of specific characteristics on the price of goods or predicting the price for a bundle of characteristics. In the second category, in particular, much research has been devoted to comparing different estimation techniques with respect to their predictive accuracy, and successively formulating recommendations on the technique to use. Unfortunately, most of the published analyses seem to suffer two main drawbacks. First, authors tend to focus on quadratic and absolute loss functions, without investigating the prediction accuracy of the chosen estimation method for other loss functions. Second, final recommendations are usually based on a direct comparison of the prediction accuracy's point estimates, completely ignoring the sample variation of such estimates.[1]

Thus, the aim of the present paper is twofold. The first purpose is to analyse how the choice of a regression method in hedonic price models will depend on different loss functions. In addition to the usual square, and absolute loss functions, a bounded loss function is also considered to compute the prediction accuracy. The second purpose is to adopt an approach that allows for statistically comparing the predictive accuracy of the considered regression methods, and to illustrate how misleading the usual comparisons based on point estimates can be. To this end, a modified version of the test proposed by Diebold and Mariano (1995), and the permutation test introduced by Konietschke and Pauly (2013) are applied to compare the predictive accuracy of non-robust, and robust regression techniques.

The paper is structured as follows. Section 5.2 introduces the employed hedonic price model, illustrates the concept of loss function, and defines the in- and out-of-sample prediction ac-

1 See Laurice and Bhattacharya (2005) and Hannonen (2008) as examples of recent publications containing these drawbacks.

curacy. The statistical tests used to compare the prediction performance of the regression methods are explained in Section 5.3. Section 5.4 describes the data, briefly illustrates the two regression methods that are compared, and shows the empirical results. Section 5.5 concludes the paper.

5.2 HEDONIC PRICE MODEL ESTIMATION AND EVALUATION

5.2.1 *Loss function and prediction accuracy*

Let the random variables P and $\mathbf{X} := (X_1, ..., X_K)$ denote the price and the K characteristics of a good, respectively. For a random sample of n independent observations

$$(P_i, \mathbf{X}_i) := (P_i, X_{i1}, ..., X_{iK}), \tag{43}$$

we assume the following additive model for the log of the price:

$$\log(P_i) = \beta_0 + X_{i1}\beta_1 + ... + X_{iK}\beta_K + \epsilon_i, \tag{44}$$

where ϵ_i represents a stochastic error term with $E(\epsilon_i|X_i) = 0$ and $V(\epsilon_i|X_i) = \sigma^2$, $\forall i$.[2]
Let $\hat{P}^r$ be the estimated price in the original scale obtained through a regression method r. In the present paper, r represents the non-robust and robust regression techniques, respectively. The main goal is to compare the distribution of the random variables $L(P, \hat{P}^r(\mathbf{X}))$, $r = 1, 2$, where $L : \mathbb{R}^2 \longrightarrow \mathbb{R}^+$ denotes the prediction loss function. Three loss functions are used in the present paper: The usual square loss function, the absolute loss function, and the bisquare loss function. For a definition of these loss functions, please refer to 5.6.
To measure the prediction accuracy, we consider the expected values $\mu_r := E(L(P, \hat{P}^r(\mathbf{X})))$, $r = 1, 2$. As noted by Hennig and Kutlukaya (2007), the choice of the central tendency indicator is arbitrary but could be related to the functional form of the loss function. Since the expected value of the loss function is not robust to extreme values, one could suggest a robust loss function to limit the influence of extreme values. This motivates the choice of the bisquare loss function.
As already mentioned, in the hedonic literature, point estimates of the expected loss are often compared, and the optimal regression method is selected according to these point-estimates.

2 Although we restrict ourselves to cross-sectional data, the illustrated procedures can easily be modified for pooled cross-sections with an heteroskedastic error term.

However, as pointed out by Diebold (2012), the fact that a single realization of the random variable μ_1 is smaller than μ_2 (or vice versa) does not guarantee that the observed inequality also holds for the population parameters. To investigate if the population parameters are statistically not equal, difference in means tests can be applied to estimators of the population parameters. The in- and out-of-sample prediction error's estimators of the expected loss are given by

$$\hat{\mu}_r^{in} = \frac{1}{n}\sum_{i=1}^{n} L(P_i, \hat{P}^r(\mathbf{X}_i)),\ r = 1, 2$$

and

$$\hat{\mu}_r^{out} = \frac{1}{n}\sum_{i=1}^{n} L(P_i, \hat{P}^{r,-s(i)}(\mathbf{X}_i)),\ r = 1, 2,$$

respectively. The expression $\hat{P}^{r,-s(i)}$ indicates that price in the original scale has been estimated without a subset of sample variables containing the i-th random variable $\mathbf{X}_i$. Both cross-validation and bootstrap techniques are available to compute $\hat{\mu}_r^{out}$ (see Hastie et al. (2003)). In the present paper, we limit ourselves to the cross-validation approach.

5.3 COMPARING PREDICTION PERFORMANCE

To gauge if the prediction accuracies of two estimation methods are statistically different for a given loss function, the following two tests are considered in the present paper:

$$H_0 : \mu_1^{in} = \mu_2^{in} \text{ against } H_1 : \mu_1^{in} \neq \mu_2^{in}$$

and

$$H_0 : \mu_1^{out} = \mu_2^{out} \text{ against } H_1 : \mu_1^{out} \neq \mu_2^{out}.$$

Let $D_i := L(P_i, \hat{P}^1(\mathbf{X}_i)) - L(P_i, \hat{P}^2(\mathbf{X}_i))$ and $\bar{D}_n := \frac{1}{n}\sum_{i=1}^{n} D_i$ denote the i-th loss differential and the mean loss differential, respectively. The observations having been assumed independent, the random variables D_i are also independent. The above hypotheses are verified by means of the following difference in means tests.

5.3.1 *Permuted t-test*

Let $L_i^r := L(P_i, \hat{P}^r(\mathbf{X}_i))$ be the i-th prediction loss for the regression method r, and $L := (L_1^1, L_2^1, ..., L_{n-1}^2, L_n^2)$ denote the 2n-

vector containing the losses of the considered regression methods. We define the vector $L^* = ((L_1^{1,*}, L_1^{2,*}), ..., (L_n^{1,*}, L_n^{2,*}))$ as a random permutation of the vector L. In a recent paper, Konietschke and Pauly (2013) define, among others, the following modified t-statistic

$$t_{KP} := \sqrt{n}\frac{\bar{D}_n^*}{V_n^*} \longrightarrow N(0,1),$$

where $D_i^* := L_i^{1,*} - L_i^{2,*}$, $i = 1, ..., n$ represent the loss differences of the permuted losses. The expressions $\bar{D}_n^* = \frac{1}{n}\sum_{i=1}^n D_i^*$ and $V_n^{*2} = \frac{1}{n-1}\sum_{i=1}^n (D_i^* - \bar{D}_n^*)^2$ simply denote the mean and variance estimators of the permuted losses.
Using a simulation study, Konietschke and Pauly (2013) show that in small samples the proposed statistic improves the power of the t-test, and the test level is nearer to the nominal one, especially for skewed distributions.

5.3.2 *Modified Diebold and Mariano test*

The following test is a modified version of the test proposed by Diebold and Mariano (1995). In this case, we do not assume that the loss differentials D_i possess the same variance. The null hypothesis of equal predictive accuracy is tested using the following modification of the usual t-statistic:

$$t_{DM} := \frac{\bar{D}_n}{\sqrt{\hat{V}(\frac{1}{n}\sum_{i=1}^n D_i)}} \longrightarrow N(0,1).$$

Due to the observations' independence, the variance $V(\frac{1}{n}\sum_{i=1}^n D_i)$ is estimated by $\frac{1}{n^2}\sum_{i=1}^n \hat{V}(D_i)$. Thus, the main issue is to consistently estimate the variances $V(D_i)$, $i = 1, ..., n$ with $V(D_i) \neq V(D_j)$ for $i \neq j$. The problem is solved by regressing the loss differentials D_i on a constant and computing the coefficient's standard error through a heteroskedastic-consistent (HC) estimator. Because of the possible presence of influential observations, we use the heteroskedastic-consistent estimator proposed by Cribari-Neto (2004).

5.4 EMPIRICAL RESULTS

5.4.1 *Data*

The data used for the analysis were kindly provided by Wüest & Partner, an international real estate consultancy firm. Transac-

tion prices of single-family dwellings and their corresponding characteristics were collected for the Swiss canton of Zurich from banks, insurances companies, and other real estate agencies. In the present study, the 411 transactions observed for the fourth quarter 2010 are considered[3].
The following characteristics were used to explain the log-price variations: age (*age*; in years), volume (*vol*; in cubic meters), surface of the land surrounding the property (*land*; in square meters), micro location within the municipality (*micro*; bad-acceptable, or *g*ood), and the macro location of the house within the canton (*macro*; center, *s*outh, or *n*orth), status (*status*; low–medium or *s*uperior), condition (*cond*; poor–reasonable or *e*xcellent), number of rooms (*rooms*), house type (*type*; semi-detached or *d*etached), and number of parking places (*garage*).

5.4.2 *Estimation comparison*

The first estimation method is the well known ordinary least squares (OLS) estimator, and the second is a robust estimation technique proposed by Maronna and Yohai (2000). In the hedonic literature, several authors have used robust regression methods to compare robust coefficients estimates to those obtained with the OLS method, finding that robust coefficients often possess a better economic interpretation[4]. Unfortunately, most of the published papers make use of M-estimators to limit the effect of influential observations (see for example Yoo (2001), Graves et al. (1988), and Janssen et al. (1984)). However, as stressed by Ellis and Morgenthaler (1992), M-estimators may have a low breakdown point even if no leverage points are present. On the contrary, the estimation method proposed by Maronna and Yohai (2000) is particularly indicated for linear hedonic regression, since it alternates M and S estimators to handle both categorical and continuous regressors, defining a high breakdown-point and computationally less expensive robust estimator. Since a Q-Q plot of the OLS residuals in equation (44) revealed a heavy-tailed distribution, a robust estimation technique seemed an appropriate way to prevent influential observations to unduly influence the model's predictions.
Table 5 contains the results of both OLS and M-S estimations. As it can be seen, the two estimation methods display simi-

3 This quarter was chosen since the hypothesis of homoskedasticity could not be rejected at the standard level of 5%.

4 For example, see Yoo (2001).

	OLS	M-S
(Intercept)	12.930***	12.842***
	(0.102)	(0.107)
age	0.000	0.001
	(0.001)	(0.001)
vol	0.001***	0.001***
	(0.000)	(0.000)
land	0.000	0.000
	(0.000)	(0.000)
status_s	0.147***	0.126**
	(0.037)	(0.039)
cond_e	0.023	0.000
	(0.039)	(0.041)
micro_g	0.148***	0.127***
	(0.034)	(0.036)
type_d	−0.058	−0.068
	(0.036)	(0.038)
rooms	0.043*	0.041*
	(0.018)	(0.019)
macro_s	0.106**	0.136**
	(0.041)	(0.043)
macro_n	−0.212***	−0.190***
	(0.044)	(0.046)
garage	0.006	0.010
	(0.018)	(0.019)
N. obs.	411	411

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5: OLS and M-S log-linear regression coefficients.

lar coefficients' values, and share exactly the same significant variables. This is not surprising, since the two estimation techniques provide consistent estimators of the population parameters. Although the coefficients are similar, the main question is whether the predictive accuracy of the two estimation techniques is the same for a given loss function.

	Square	Absolute	Bisquare
OLS	0.233	0.279	0.882
Robust	0.273	0.283	0.866

Table 6: In-sample mean losses.

5.4.3 *In-sample prediction accuracy*

The in-sample prediction performance of the two estimation methods was compared for the square, absolute, and bisquare loss functions. Once the log-linear regression model was estimated, the predictions in the original scale were obtained using the smearing estimator proposed by Duan (1983). Duan's smearing estimator guarantees that predictions in the original scale are unbiased for a general distribution of the regression model's error term.
Table 6 shows the mean losses of the two estimation techniques for the considered loss functions. Apparently, the OLS method performs much better than the robust one for a square loss function. This result could be interpreted as a consequence of the fact that the OLS estimator minimizes the sum of square residuals, thereby providing the smallest expected square loss. For the absolute loss function, the two techniques seem to possess similar expected losses. Finally, the M-S estimator seems to possess a slightly lower expected bisquare loss. This result could also be expected, since the considered robust estimator alternates estimators that minimize robust loss functions, thus providing a better performance for a bounded prediction loss function. Except for the bisquare loss function, the OLS estimation technique should, therefore, be preferred to its robust counterpart. However, as illustrated in Table 7, the above interpretation is not correct. Both the Diebold-Mariano (D-M) and Konietschke-Pauly (K-P) t-tests reveal that the prediction accuracy of the

	Square	Absolute	Bisquare
D-M	0.201	0.446	0.042
K-P	0.392	0.469	0.038

Table 7: In-sample difference in means: p-values of equality of means tests.

Fold	Square	Absolute	Bisquare
1	0.486	0.174	0.787
2	0.369	0.375	0.866
3	0.101	0.393	0.868
4	0.960	0.359	0.814
5	0.774	0.689	0.225
All	0.501	0.455	0.441

Table 8: Out-of-sample difference in means: p-values of the K-P test.

OLS estimator is not statistically different from that of the robust M-S estimator for the square and absolute loss functions. For the bisquare loss function, however, the mean losses are significantly different at the standard level of 5%. According to the computed p-values, the M-S estimator should, therefore, be preferred to the OLS estimator, since it performs at least as good as the OLS estimator for the considered loss functions.

5.4.4 *Out-of-sample prediction accuracy*

In this section, the out-of-sample prediction performance of the OLS and M-S estimators is analyzed. A 5-fold cross-validation was carried out to compute the out-of-sample prediction losses of each regression technique. To effectuate a direct comparison of the prediction accuracy, the OLS and M-S estimators were computed using the same set of randomly generated folds, and the out-of-sample losses estimated on the same left-out fold. To assess how the expected out-of-sample prediction was influenced by an estimation technique in a given fold, the prediction performance was compared both within each fold and globally using the K-P test[5].
Table 8 shows the obtained results. As can be seen, none of the considered techniques seems to possess a better prediction performance for the three loss functions. This result is in contrast to what was obtained for the in-sample prediction performance, where the robust method dominated the OLS estimator for the bisquare loss function. It turns out, however, that the out-of-sample prediction performance of the robust method is strongly dependent on the set of randomly generated folds. It

5 The D-M test provided similar results.

seems that the more the influential observations are equally distributed within each fold, the more similar are the OLS and M-S prediction performance. On the contrary, if the influential observations are concentrated in a small number of folds, the robust estimation technique performs better than the OLS technique in these folds, causing its global out-of-sample expected loss to be statistically lower than that of the OLS estimator. Increasing the number of folds worsens the results' stability for the given number of observations. Although the expected out-of-sample losses of the two regression methods do not seem to be statistically different, the computed p-values must be cautiously interpreted.

5.5 CONCLUSIONS

In the present paper, a systematic approach to compare the predictive accuracy of two regression methods was presented. In particular, using cross-sectional data, the prediction performance of the usual OLS estimator was compared to the prediction accuracy of the robust M-S estimator proposed by Maronna and Yohai (2000). The prediction accuracy of the two regression methods was compared for the square, absolute, and bisquare loss functions. The Diebold-Mariano and Konietschke-Pauly tests were finally applied to assess if the expected in- and out-of-sample losses were statistically different among the regression methods.

A rough comparison of the expected loss point estimates, as often effectuated in the hedonic literature, was shown to be misleading, since it does not account for the sample variation of these point estimates. Moreover, the out-of-sample predictive accuracy may be strongly dependent on the resampling technique used to compute out-of-sample predictions. Thus, great care is needed in the interpretation of the out-of-sample prediction accuracy. Based on the introduced methodology, the M-S estimator was shown to perform as well as the OLS estimator for the square and absolute loss functions. Interestingly, despite the two regression methods displaying similar coefficient values, the M-S estimator performed significantly better for the bisquare loss function.

In the present paper, we concentrated on comparing the predictive accuracy of two regression methods. The same methodology could also be applied to compare the prediction performance of a given estimation technique for different loss func-

tions. Additionally, asymmetric loss functions could be employed to determine the choice of regression method to use. This is an area for further research.

5.6 APPENDIX: LOSS FUNCTIONS

Let P and $\hat{P}$ denote the model's dependent variable and its predicted value, respectively. The following loss functions have been considered in the present analysis:

- Square loss function:

$$L_{Square} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}^+$$
$$(P, \hat{P}) \longmapsto \left(\frac{P - \hat{P}}{P} \right)^2 .$$

- Absolute loss function:

$$L_{Abs} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}^+$$
$$(P, \hat{P}) \longmapsto \left| \frac{P - \hat{P}}{P} \right| .$$

- Bisquare loss function:

$$L_{Bisq} : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}^+$$
$$(P, \hat{P}) \longmapsto \begin{cases} 1 - \left(1 - \left(\frac{1}{k} \frac{P - \hat{P}}{P} \right)^2 \right)^3 & \text{if } \left| \frac{P - \hat{P}}{P} \right| \leqslant k \\ 1 & \text{if } \left| \frac{P - \hat{P}}{P} \right| > k, \end{cases}$$

  where the k represents a parameter defining the function's shape.

The three loss functions are represented in Figure 3.
Three remarks are necessary to fully understand the used loss functions. First, the above loss functions depend on the relative residuals $e_r := (P - \hat{P})/P$. This seems more reasonable in the hedonic price context than to consider loss functions based on usual prediction errors. It appears unrealistic to assume that the loss caused by the prediction error $P - \hat{P}$ does not depend on the relative value of P: A prediction error of 100'000 CHF should cause a greater loss for a house worth 500'000 CHF than for a house worth 1'000'000 CHF. Second, only symmetric loss functions have been used, i.e., $L(e_r) := L(P, \hat{P}) = L(-e_r)$. This implies that overestimating the price by a given amount causes exactly the same loss as underestimating the price by the same amount. This assumption could not be true and depends on the purpose for which price prediction is conducted.

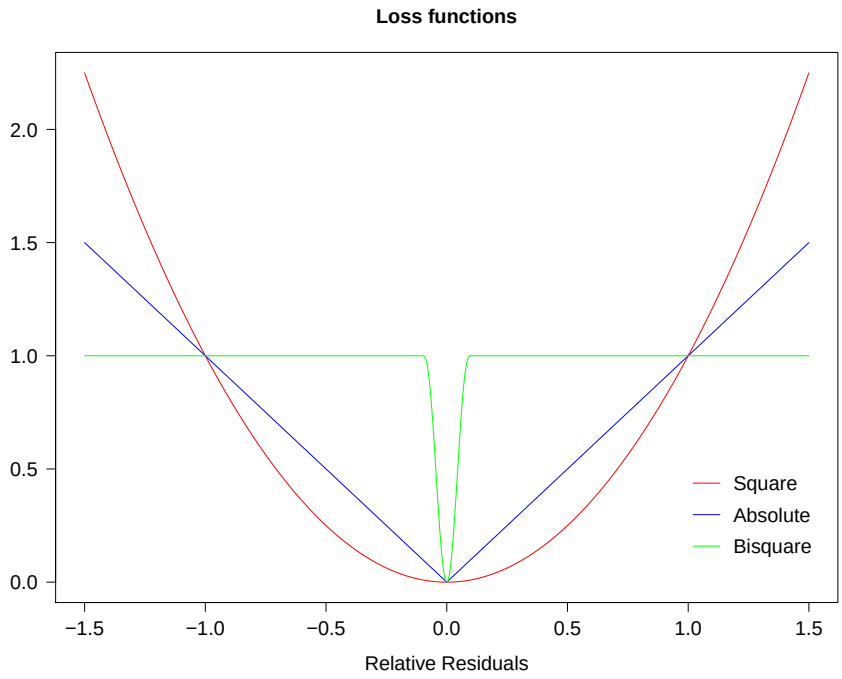

Figure 3: Loss functions.

Consider the point of view of a person who owns a house. If the aim is to predict the price for sale purposes, overestimating the price might cause a lesser loss than underestimating the price by a given amount for this person. On the contrary, the same person might experience a greater loss caused by overestimating the house price when tax valuation is considered. Third, as stressed by Hennig and Kutlukaya (2007), the shape parameter k should not depend on the data, i.e. it should be determined a priori. For the present analysis, a value of $k = 10\%$ has been chosen. The shape parameter should be interpreted as follows. If a prediction error exceeds 10% of the property value, the loss caused does not depend on the magnitude of the prediction error. This could be the case, for example, if the person for which the home price is predicted does not take into account the prediction's validity for relative errors greater than 10%. The shape parameter, therefore, defines a kind of tolerance threshold for the person for whom the prediction is conducted.

Part III

FINAL REMARKS

6

CONCLUSIONS AND FURTHER RESEARCH

The hedonic domain represents a rich research field in which a wide range of statistical tools can be used to provide an answer to valuing goods with non-constant quality. The present thesis tries to shed light on several aspects of hedonic estimation in the housing field in general, and, in particular, on the estimation of hedonic price indices. Using theoretical and empirical arguments, several results have been achieved regarding these applications. Many problems, however, remain unsolved, and represent an area of further research.

The price index problem in the hedonic context, in particular, is far from being solved. It could be interesting, for example, to empirically investigate how nonlinear hedonic functions influence the price index problem, and establish if asymptotically the price index formulae are statistically different. This could be achieved by computing, for different hedonic functions, a given price index formula. The equality of the price indices could then be tested, and conclusions drawn regarding the most stable price index formula.

Concerning the variable selection techniques in the presence of multicollinearity, the approach using several competing models seems to be promising. Comparing the proposed selection algorithm to a selection algorithm using a Bayesian approach could be interesting. Moreover, the performance of the proposed selection algorithm was simulated under normally distributed errors. Its performance for non-spherical errors is unknown and requires further investigation. In particular, a simulation using heavy-tailed distribution, a class of distribution often observed in practice, could be captivating.

The proposed methodology for assessing the prediction performance of hedonic models under general loss function, could be used to compute hedonic price indices. According to the aim of the price index, a loss function could be modeled, and a corresponding estimation technique chosen. The price index would finally be computed according to this estimation technique that minimized the expected loss. Also in this case, investigating how the computed price index differs for different loss functions could be interesting.

In conclusion, many questions in the hedonic domain remain unanswered, and further research is needed to improve our knowledge in this field.

BIBLIOGRAPHY

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proceedings of 2nd International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, 1973.

Bert Balk. Price indexes for elementary aggregates: The sampling approach. *Journal of Official Statistics*, 21(4):675–99, 2005.

Silvia Banfi, Massimo Filippini, and Andrea Horehájová. Valuation of environmental goods in profit and non-profit housing sectors: Evidence from the rental market in the city of zurich. *Swiss Journal of Economics and Statistics (SJES)*, 144 (IV):631–654, 2008.

Andrea Baranzini and José Ramirez. Paying for quietness: The impact of noise on geneva rents. *Urban Studies (Routledge)*, 42 (4):633 – 646, 2005.

Andrea Baranzini and Caroline Schaerer. A sight for sore eyes: Assessing the value of view and land use in the housing market. *Journal of Housing Economics*, 20(3):191–199, 2011.

Andrea Baranzini, José Ramirez, Caroline Schaerer, and Philippe Thalmann. Introduction to this volume: Applying hedonics in the swiss housing markets. *Swiss Journal of Economics and Statistics (SJES)*, 144(IV):543–559, 2008a.

Andrea Baranzini, Caroline Schaerer, José Ramirez, and Philippe Thalmann. Using the hedonic approach to value natural land uses in an urban area: An application to Geneva and Zurich. *Economie publique / Public Economics*, 20(2007/1): 147–167, 2008b.

Andrea Baranzini, Caroline Schaerer, José Ramirez, and Philippe Thalmann. Do foreigners pay higher rents for the same quality of housing in geneva and zurich? *Swiss Journal of Economics and Statistics (SJES)*, 144(IV):703–730, 2008c.

Michael Beer. *Hedonic Elementary Price Indices: Axiomatic Foundation and Estimation Techniques*. PhD thesis, University of Fribourg, 2006.

Michael Beer. *Hedonic Elementary Price Indices: Axiomatic Foundation and Estimation Techniques*. PhD thesis, University of Fribourg Switzerland, 2007a.

Michael Beer. Bootstrapping a hedonic price index: Experience from used cars data. *AStA Advances in Statistical Analysis*, 91 (1):77–92, 2007b.

Andre Bender, Gacem Brahim, and Martin Hoesli. Construction d'indices immobiliers selon l'approche he?doniste. *Finanzmarkt Und Portfolio Management*, (8):522–534, 1994.

Franziska Bignasca, Roswitha Kruck, and Rico Maggi. Immobilienmarkt zürich - immobilienpreise und bauinvestitionen unter der lupe. Technical report, Züricher Kantonalbank: Wirtschaft und Gesellschaft, 1996.

Steven Bourassa, Martin Hoesli, Donato Scognamiglio, and Philippe Sormani. Constant-quality house price indexes for switzerland. Swiss Finance Institute Research Paper Series 08-10, Swiss Finance Institute, 2008.

Steven Bourassa, Martin Hoesli, Donato Scognamiglio, and Sumei Zhang. Land leverage and house prices. *Regional Science and Urban Economics*, 41(2):134–144, 2011.

George Box and David Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):pp. 211–252, 1964.

Hans Wolfgang Brachinger. Statistical theory of hedonic price indices. DQE Working Paper 1, Department of Quantitative Economics, University of Fribourg Switzerland, 2002.

Hans Wolfgang Brachinger and Michael Beer. The Econometric Foundations of Hedonic Elementary Price Indices. DQE Working Papers 12, Department of Quantitative Economics, University of Freiburg/Fribourg Switzerland, 2009.

Hans Wolfgang Brachinger, Michael Beer, and Olivier Schöni. The Econometric Foundations of Hedonic Elementary Price Indices. Working paper, Department of Quantitative Economics, University of Freiburg/Fribourg Switzerland, 2012.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim Buhmann. The balanced accuracy and its posterior distribution. In *Proceedings of the 2010 20th International*

Conference on Pattern Recognition, ICPR '10, pages 3121–3124, Washington, DC, USA, 2010. IEEE Computer Society.

Kenneth Burnham and David Anderson. Multimodel Inference. *Sociological Methods & Research*, 33(2):261–304, November 2004.

Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1-2):103–112, July 2005.

Andrew Court. Hedonic price indexes with automotive examples. In *The Dynamics of Automobile Demand*, pages 99–117, New York, 1939. General Motors Corporation.

Francisco Cribari-Neto. Asymptotic inference under heteroskedasticity of unknown form. *Computational Statistics & Data Analysis*, 45(2):215–233, 2004.

Bruce Curry, Peter Morgan, and Mick Silver. Hedonic regressions: Mis-specification and neural networks. *Applied Economics*, 33(5):659–671, 2001.

Anirban DasGupta. *Probability for Statistics and Machine Learning*. Springer Texts in Statistics, 2011.

Russell Davidson and Emmanuel Flachaire. The wild bootstrap, tamed at last. Econometric society world congress 2000 contributed papers, Econometric Society, 2000.

Mark Dickie, Charles Delorme, Jr, and Jeffrey Humphreys. Hedonic prices, goods-specific effects and functional form: inferences from cross-section time series data. *Applied Economics*, 29(2):239–49, 1997.

Francis Diebold. Comparing predictive accuracy, twenty years later: A personal perspective on the use and abuse of diebold-mariano tests. Working paper, National Bureau of Economic Research, 2012.

Francis Diebold and Roberto Mariano. Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–263, 1995.

Erwin Diewert, Saeed Heravi, and Mick Silver. Hedonic imputation versus time dummy hedonic indexes. In Erwin Diewert,

John Greenlees, and Charles Hulten, editors, *Price Index Concepts and Measurement*, chapter 4, pages 161–196. University of Chicago Press, 2009. ISBN 0-226-14855-6.

Dragana Djurdjevic, Christine Eugster, and Ronny Haase. Estimation of hedonic models using a multilevel approach: An application for the swiss rental market. *Swiss Journal of Economics and Statistics (SJES)*, 144(IV):679–701, 2008.

Alan Dorfman, Sylvia Leaver, and Janice Lent. Some observations on price index estimators. Statistical Policy Working Paper 29, Bureau of Labor Statistics, 1999.

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

John Eatwell, Murray Milgate, and Peter Newman, editors. *The New Palgrave: A Dictionary of Economics*. The Macmillan Press Limited, London, 1987.

Wolfgang Eichhorn. What is an economic index? an attempt of an answer. In Wolfgang Eichhorn, R. Henn, O. Opitz, and R. W. Shephard, editors, *Theory and Applications of Economic Indices*, pages 3–42. Physica-Verlag, Würzburg, 1978.

Wolfgang Eichhorn and Joachim Voeller. *Theory of the Price Index*. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, Berlin, 1976.

Steven Ellis and Stephan Morgenthaler. Leverage and breakdown in l1 regression. *Journal of the American Statistical Association*, 87(417):pp. 143–148, 1992.

Stefan Fahrländer. Hedonische handänderungspreisindizes: Konzeption und methodik. Technical report, Wüest & Partner, 2001a.

Stefan Fahrländer. Performance von wohneigentum. Technical report, Wüest & Partner, 2001b.

Stefan Fahrländer. Semiparametric construction of spatial generalized hedonic models for private properties. *Swiss Journal of Economics and Statistics (SJES)*, 142(IV):501?528, 2006.

Stefan Fahrländer. Indirect construction of hedonic price indexes for private properties. *Swiss Journal of Economics and Statistics (SJES)*, 144(IV):607–630, 2008.

Mike Fletcher, Paul Gallimore, and Jean Mangan. Heteroscedasticity in hedonic house price models. *Journal of Property Research*, 17(2):37–41, 2000.

Allen Goodman and Thomas Thibodeau. Age-related heteroskedasticity in hedonic house price equations. *Journal of Housing Research*, 6(1):25–42, 1995.

Allen Goodman and Thomas Thibodeau. Dwelling-age-related heteroskedasticity in hedonic house price equations: An extension. *Journal of Housing Research*, 8(2):299–317, 1997.

Phil Graves, James Murdoch, Mark Thayer, and Don Waldman. The robustness of hedonic price estimation: Urban air quality. *Land Economics*, 64(3):pp. 220–233, 1988.

William Greene. *Econometric Analysis - 7 Edition*. Pearson Education, 2011.

Zvi Griliches. Hedonic price indexes for automobiles: An econometric analysis of quality change. In Zvi Griliches, editor, *Price Indexes and Quality Change*, pages 55–87. Harvard University Press, Cambridge, 1971.

Pascal Grosclaude and Nils Soguel. Coûts externes du trafic routier: évaluation en milieu urbain. *Swiss Journal of Economics and Statistics (SJES)*, 128(III):453–469, 1992.

Marko Hannonen. Predicting urban land prices: A comparison of four approaches. *International Journal of Strategic Property Management*, 12(4):217–236, 2008.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, corrected edition, 2003.

Christian Hennig and Mahmut Kutlukaya. Some thoughts about the design of loss functions. *REVSTAT-Statistical Journal*, 5(1):19–39, 2007.

Robert Hill. Hedonic Price Indexes for Housing. OECD Statistics Working Papers 2011/01, OECD Publishing, 2011.

Robert Hill. Hedonic price indexes for residential housing: A survey, evaluation and taxonomy. *Journal of Economic Surveys*, 2012.

Robert Hill and Daniel Melser. Hedonic Imputation and the Price Index Problem: an Application To Housing. *Economic Inquiry*, 46(4):593–609, 2008.

Martin Hoesli, Philippe Favarger, and Carmelo Giaccotto. Real estate price indices and performance: The case of geneva. *Swiss Journal of Economics and Statistics (SJES)*, 133(I):29–48, 1997a.

Martin Hoesli, Carmelo Giaccotto, and Philippe Favarger. Three new real estate price indices for geneva, switzerland. *The Journal of Real Estate Finance and Economics*, 15(1):93–109, 1997b.

ILO, IMF, OECD, UNECE, Eurostat, and The World Bank, editors. *Consumer Price Index Manual: Theory and Practice*. International Labour Office, Geneva, 2004.

ILO, IMF, OECD, UNECE, Eurostat, and The World Bank, editors. *Consumer Price Index Manual: Theory and Practice*. International Labour Office, Geneva, 2010. Partially revised version published online [accessed 29 January 2011].

Rolf Iten and Markus Maibach. Externe kosten durch verkehrslärm in stadt und agglomeration zürich. *Swiss Journal of Economics and Statistics (SJES)*, 128(I):51–68, 1992.

Christian Janssen, Bo Söderberg, and Julie Zhou. Robust estimation of hedonic models of price and income for investment property. *Journal of Property Investment & Finance*, 19(4):342–360, 1984.

Frank Konietschke and Markus Pauly. Bootstrapping and permuting paired t-test type statistics. *Statistics and Computing*, pages 1–14, 2013.

Kelvin Lancaster. A new approach to consumer theory. *The Journal of Political Economy*, 74(2):132–57, 1966.

Kelvin Lancaster. *Consumer Demand: A New Approach*. Number 5 in Columbia Studies in Economics. Columbia University Press, New York, 1971. ISBN 0-231-03357-5.

Manuel Landajo, Celia Bilbao, and Amelia Bilbao. Nonparametric neural network modeling of hedonic prices in the housing market. *Empirical Economics*, 42(3):987–1009, 2012. ISSN 0377-7332.

Jennifer Laurice and Radha Bhattacharya. Prediction performance of a hedonic pricing model for housing. *The Appraisal Journal*, 73(2):198–209, 2005.

Scott Long and Laurie Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

Ricardo Maronna and Victor Yohai. Robust regression with both continuous and categorical predictors. *Journal of Statistical Planning and Inference*, 89(1-2):197–214, 2000.

Raimond Maurer, Martin Pitzer, and Steffen Sebastian. Hedonic price indices for the paris housing market. *Allgemeines Statistisches Archiv*, 88(3):303–326, 2004.

Murray Milgate. Goods and commodities. In Eatwell et al. (1987), pages 546–9.

Jon Nelson. Valuing rural recreation amenities: Hedonic prices for vacation rental houses at deep creek lake, maryland. *Agricultural and Resource Economics Review*, 39(3), 2010.

OECD. Handbook on residential property price indices. Technical report, 2013.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012.

Peter Riedi. *Die Entwicklung der Bodenpreise im Kanton Tessin in den Jahren 1978 bis 1989 und deren Beeinflussungsfaktoren*. PhD thesis, University of Bern, 1992.

Sherwin Rosen. Hedonic prices and implicit markets: Product differentiation in pure competition. *The Journal of Political Economy*, 82(1):34–55, 1974.

Marco Salvi. Spatial estimation of the impact of airport noise on residential housing prices. *Swiss Journal of Economics and Statistics (SJES)*, 144(IV):577–606, 2008.

Marco Salvi, Patrik Schellenbauer, and Hansjörg Schmidt. Preise, mieten und renditen - der immobilienmarkt transparent gemacht. *Zurcher Kantonalbank*, 2004.

Gideon Schwarz. Estimating the dimensions of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Donato Scognamiglio. *Methoden zur Immobilienbewertung im Vergleich: eine empirische Untersuchung für die Schweiz*. PhD thesis, University of Bern, 2000.

Mick Silver and Saeed Heravi. The measurement of quality-adjusted price changes. In Robert Feenstra and Matthew Shapiro, editors, *Scanner Data and Price Indexes*, volume 64 of *Studies in Income and Wealth*, pages 277–316, Chicago, 2003. The University of Chicago Press.

Mick Silver and Saeed Heravi. Why elementary price index number formulas differ: Evidence on price dispersion. *Journal of Econometrics*, 140(2):874–83, 2007.

Nils Soguel, Marc-Jean Martin, and Alexandre Tangerini. The impact of housing market segmentation between tourists and residents on the hedonic price for landscape quality. *Swiss Journal of Economics and Statistics (SJES)*, 144(IV):655–678, 2008.

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

Simon Stevenson. New empirical evidence on heteroscedasticity in hedonic housing models. *Journal of Housing Economics*, 13(2):136–53, 2004b.

Philippe Thalmann. Explication empirique des loyers lausannois. *Swiss Journal of Economics and Statistics (SJES)*, 123(I): 47–70, 1987.

Jack Triplett. Hedonic functions and hedonic indexes. In Eatwell et al. (1987), pages 630–4.

Jack Triplett. Handbook on hedonic indexes and quality adjustments in price indexes: Special application to information technology products. Working Paper 2004/9, OECD Directorate for Science, Technology and Industry, Paris, 2004.

United Nations, editor. *System of National Accounts 1993*, chapter XVI. Price and Volume Measures. United Nations, 1993.

Eric White and Larry Leefers. Influence of natural amenities on residential property values in a rural setting. *Society & Natural Resources*, 20(7):659–667, 2007.

Seung-Hoon Yoo. A robust estimation of hedonic price models: least absolute deviations estimation. *Applied Economics Letters*, 8(1):55–58, 2001.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Final Version as of March 20, 2014 (`classicthesis` version 4.1).