Hedonic Price Index Number for Blocks of Flats and Terraced Houses in Finland

Suoperä, Antti, Statistics Finland
Luomaranta, Henri, Statistics Finland
Kaila, Johanna, Statistics Finland

Abstract

Statistics Finland has abandoned a standard Griliches-type time-dummy hedonic approach in the beginning of 2000 for several reasons. First of all, the quality adjustment is not easily interpretable, which is essential in practical index number compilation. Secondly, the empirical results show that aggregation method has consistency problems and may cause severe bias for price change estimates. Finally, it is not rooted in the traditional index number theory. To overcome these problems, Statistics Finland developed two new hedonic approaches - first for old blocks of flats and terraced houses and then for rents of office and shop premises. Both analyses are based on well-known decomposition introduced in Oaxaca (1973). Statistics Finland was the first to apply this decomposition in the hedonic price index compilation. The idea is to explicitly decompose the average price changes into quality correction and quality adjusted price change components, both for unweighted geometric average and weighted arithmetic average cases.

In this study, we provide a synthesis of these two hedonic techniques by discussing transparent economic, mathematical and statistical explanations for the used methods. The process of index compilation is not straightforward and requires economic thinking in producing regional partition of transactions, statistical inference by using the fixed effects model, and finally, the entire process relies on standard index number theory in the price aggregation step. We aim at providing an accessible blueprint and explanation of the methods for official statistics practitioners. The empirical tests rely on high-quality quarterly level register on transactions containing prices, quantities, values and basic quality characteristics from 2015/I-2020/IV.

Keywords
Partition, Unit Value, Logarithmic Representations, Index Number Formulas, Hedonic Method, FE-Model, OLS Method, Price Aggregation, Oaxaca Decomposition, Logarithmic mean.

1 Introduction

The traditional index number theory is based on bilateral price-links where price changes are measured from comparable commodities. In this spirit Bailey, Muth and Nourse (1963) develop a repeat-sales model (see, Case and Shiller,1989; Quigley, 1995) using a model based (or the stochastic) approach to index numbers in which uncertainty and statistical ideas play a central role. These repeat-sales models are problematic, because they can only capture a tiny fraction of the data in Finland because each transacted dwelling appears rarely more than once in the data in a short time span.

Another well-known model-based approach is Griliches (1973) time-dummy hedonic method, which is able to cover the entire data by resolving the comparability issue by applying quality adjusting. The method uses standard regression analysis and in statistical point of view the method seems simple and understandable. Unfortunately, the method suffers from several problems in index number calculations. First, the analysis of quality correction is not transparently presented, so that the quality adjusted price changes are difficult to interpret. Second, the method agrees with the index number theory only in very specific conditions. Third, the price change estimates of the Griliches-type model (see for example WTPD-model, Diewert and Fox, 2018, pp.15) may include also some quantity change caused by asymmetric weighting, leading to serious biases in some situations (see Suoperä, Nieminen, Montonen and Markkanen, 2021, pp. 12-14). Fourth, applications for several aggregation levels are problematic and probably impossible to apply consistently.

This work builds on two earlier papers (Koev, 2003; Suoperä, 2006; see also Vartia, Suoperä & Vuorio, 2021; Suoperä & Auno, 2021) which address most of above issues based on hedonic approach to index numbers. The main idea is, that because effective matched pairs method or bilateral price-linking is not possible, the price-linking should be done for some, coarse but homogeneous, grouping of observations. In statistical terms, this means partitioning of observations into strata, which in the case of house prices, is naturally based on regional classification. Unfortunately, observations in strata are not comparable in quality over time and the quality adjusting remains a problem in stratum level aggregates - whatever they are. According Koev (2003) and Suoperä (2006), a solution to the problem is two-fold: First, the partition is combined with regression analysis into well-known Fixed Effect model (Hsiao, 1986, s.29-32) and second, the quality adjusting is performed using decomposition introduced by Oaxaca (1973). Now, instead of bilateral price-linking of observations, we may make price-linking for some stratum aggregate, say, for example for unweighted geometric average or for weighted arithmetic average. The decomposition split the true average price change consistently into two parts: quality change(s) and quality adjusted price changes for any stratum in question. The decomposition makes it possible to apply index number theory for stratum level aggregates of the decomposition similarly as in traditional number theory. We analyze two stratum aggregates and their decompositions – traditional unweighted geometric average and rarely used weighted arithmetic average – and apply classical index number theory to them. All analyses are done using logarithmic representations. Practically, we follow Koev (2003); Suoperä (2006); Suoperä, Nieminen, Montonen and Markkanen (2021); Vartia, Suoperä & Vuorio (2021) and Suoperä & Auno (2021) and pick up the most suitable methods to construct hedonic price index numbers for block of flats and terraced houses in Finland. We perform our analysis of index numbers using several basic (Laspeyres (L), log-Laspeyres (l), Log-Paasche (p), Paasche (P)) and excellent index number formulas (Törnqvist (T), Montgomery-Vartia (MV), Sato-Vartia (SV), Fisher (F)).

The structure of the study is as follows. In chapter 2 we present the data, basic concepts and notations. In chapter 3 we present two nested partitions and combine them with heterogeneously behaving cross-sectional regressions. In chapter 4 we derive stratum aggregates and their Oaxaca decompositions. In chapter 5 we apply index number methods to our stratum aggregates and chapter 6 concludes.

2 Data, Basic Concepts and Notation

2.1 Data

Our data is derived from a register maintained by the tax authority. It covers all transactions of dwellings for blocks of flats and terraced houses in Finland. Statistics Finland have applied further treatment to the data by merging dwelling specific information (i.e., quality characteristics) from other registers. In this study we analyze the data quarterly from 2015/I to 2020/IV. The data includes about 17 000 and 35 000 transactions for terraced houses and for blocks of flats respectively. Both dwelling-types are analyzed separately.

2.2 Basic Concepts

Price is defined as dwelling specific *unit value* measuring *price per square meter* of a dwelling. In this study, the unit prices are in logarithmic scale. All other variables are measured by their typical units of measurement - size of dwelling in square meters, distance of dwelling from center of municipal services in minutes and age of dwelling in years. In short, this means that our price model is specified as semilogarithmic.

2.3 Notation

The notations in this study are two-fold. First, in observation level we use typical econometric notation, because we use model-based price analysis. Aggregation of variables (i.e., dependent and independent) from observations into strata (i.e., into index commodities or stratum aggregates) connect notations also into traditional notations of index number theory. The most important concepts are:

Observation level:
Commodities: $a_1, a_2, \dots, a_{n_t}$ are transacted dwellings in period *t* (blocks of flats or terraced houses).
Time periods: *t* = 0, 1, 2, … are the compared quarters.
Quantity: $q_i^t = q_{it}$ is the size of dwelling of a_i as square meters in period *t.*
Unit value or unit price: $p_i^t = v_i^t / q_i^t$ or $p_{it} = v_{it}/q_{it}$ is the price of dwelling a_i per square meter in period *t*
Value: $v_i^t = v_{it} = q_{it}p_{it}$ is the value of dwelling a_i in period *t.*
Total value: $V^t = \sum_i v_i^t = \sum_i v_{it}$ is the total value of all dwellings in period *t.*
Total quantity: $Q^t = \sum_i q_i^t = \sum_i q_{it}$ is the total quantity of all dwellings in period *t.*
Explanatory variables in regressions: $\boldsymbol{x}_{it} = (x_{it1} \dots x_{itk})'$ is a *k*-vector of observed characteristics in period *t.*

Stratum level (i.e., elementary aggregates, for example conditional averages):
Price relatives: $\bar{p}_k^{t/0} = \bar{p}_{kt}/\bar{p}_{k0}$ is the price relative of averaged unit prices for stratum *k* from period 0 to *t.*
Quantity relatives: $q_k^{t/0} = q_{kt}/q_{k0}$ is the quantity relative for stratum *k* from period 0 to *t.*
Value relatives: $v_k^{t/0} = v_{kt}/v_{k0}$ is the value relative for stratum *k* from period 0 to *t.*
Value shares: $w_{kt} = v_{kt}/\sum_k v_{kt}$ is the value share for stratum *k* in period *t.*
Explanatory variables in regressions: $\bar{\boldsymbol{x}}_{kt} = (\bar{x}_{t1} \dots \bar{x}_{tk})'$ is a *k*-vector of averaged characteristics for stratum *k* in period *t.*

The averaged variables will be defined more specifically when different aggregation rules are used.

3 The Regression Analysis Stage

When traditional price-linking (i.e., methods of bilateral price-links and repeat-sales model or matched pairs) is not available for commodities comparable in quality a partition of statistical units is necessary. Partition means for most statisticians classification of statistical units into most ‘homogenous’ disjoint stratums. For theorists, this may seem as ‘a piece of cake’, but in empirical analyzes definition of partition is very complicated –

homogeneous groupings is not easy to come by and may cause serious problems when price change estimates are calculated for stratums using a stochastic approach. In this study we test two competing partitions that are both based on regions and room numbers. The second partition is much more detailed than the first one, besides regions and sub-areas it includes the most important postal code areas and smaller municipalities. Similarly, as in our earlier studies (Suoperä and Auno, 2021; Suoperä, Nieminen, Montonen and Markkanen, 2021), regression analysis combined with partition may lead to price change estimates for stratums (i.e., commodity groups) that are severely biased.

We proceed similarly as in Suoperä and Vartia (2011) – in first stage we make partitions of transacted dwellings and then apply regression analysis in each partition. These two stages are closely related and here we combine them into fixed-effects dummy-variable approach (Hsiao, 1986, s.29-32). We show that regression analysis combined with the partition is operational especially in construction of hedonic index numbers (Koev, 2003; Suoperä, 2003, 2004, 2007, 2009, 2010).

3.1 Partition of Transacted Dwellings

We define separate partitions for block of flats and terraced houses. Transacted dwellings $A=\left\{a_1,a_2,a_3,...,a_n\right\}$ are stratified into strata A_k where sub-index k = *1,...,K* represents the stratum. Subpopulations A_k of dwellings are separate and exclude each other, that is, $A_k \cap A_{k'} = \emptyset, \quad \forall k \neq k'$ and A = $\bigcup_{k=1}^{K} A_k$. This is the simplest mathematical definition of partition. Its empirical counterpart follows Koev (2003, Appendix, Regional stratification, p. 54). Location, type of building and number of rooms are the most fundamental characteristics of the dwelling and prices vary according to these characteristics the most (Koev, 2003, p.21). Location of dwellings are based on regional stratification and within each region the dwellings are divided by type and number of rooms as follows:

Apartments in blocks of flats			Apartments in terraced houses	
1 room	2 rooms	at least 3 rooms	1 or 2 rooms	at least 3 rooms

We define two competing partitions for both types of dwellings. For terraced houses the first partition includes 110 stratums including most important regions and their sub-areas for the apartment-types shown above. The second partition is derived applying additional stratification to the first one by including the most important postal areas into it. So, we follow the idea of Koev (2003, pp.31) and apply postal areal indicators for the municipalities, which are separately examined and mere municipal indicators for the rest of the regions. The second partition for terraced houses includes about 1350 stratums. For blocks of flats two partitions are done similarly – first partition includes 165 and second one about 1730 strata.

Two nested competing partitions are applied first for statistical inference of price model (see Suoperä & Vartia, 2011, p. 21, Table 5.2) and second to test whether the first partition is detailed enough or is there a need for a more detailed partition. This will be done following Suoperä and Auno (2021) and Suoperä, Nieminen, Montonen and Markkanen (2021). In the index number chapter these two partitions are used as classification index, which measures the price change for unweighted geometric and weighted arithmetic average prices.

3.2 The Price Model for Heterogeneously Behaving Cross-sections

Because of algebraic properties of the Gauss LS-regression (i.e., least squares) our focus is analyzing vector of averages for stratum A_k, say $(\bar{p}_{k0}, \bar{p}_{kt}, \bar{\boldsymbol{x}}_{k0}, \bar{\boldsymbol{x}}_{kt})$, where $\bar{p}$-variables are some average prices and $\bar{\boldsymbol{x}}$-vectors corresponding averages of quality characteristics of dwellings for stratum A_k in time periods 0, t. When $\bar{\boldsymbol{x}}_{kt} -$

$\bar{\boldsymbol{x}}_{k0} \approx \mathbf{0}$ quality adjusting is unnecessary, otherwise quality adjustment is needed. For quality adjustment we use regression analysis that is combined with two partitions defined in chapter 3.1.

We define 34 separate estimation areas for terraced houses and block of flats. 28 of these estimation areas are largest municipals and 6 of them are based on Nuts2 or larger areas. For these estimation areas (i.e., separate price models) we apply two nested partitions: First our *basic regional partition* (terraced houses 110 and block of flats 165 stratums) and second, our basic regional partition is stratified by additional stratification based on postal areas or smaller municipals. Simply put, our two partitions are *nested* or *hierarchical* together. Our problem is to find a proper partition of transacted dwellings using principles of statistical inference and to make sure that the quality differences are controlled properly.

Next, we define price model for some estimation area, say for Helsinki – all other estimation areas (34 areas) are analyzed analogously. The price model is specified as semilogarithmic regression model, which is linear with respect to parameters, that is (see Hsiao, 1986, s.29-32)

$$(1) \qquad \log(p_{it}) = \beta_{01t} + \cdots + \beta_{0k_1t} + \boldsymbol{x}'_{it}\boldsymbol{\beta}_t + \varepsilon_{it},$$

where $\log(p_{it})$ presents dwelling specific logarithmic unit value (i.e., unit price) per square meter in estimation area Helsinki in period *t*. The k-dimensional vector of unknown parameters $\boldsymbol{\beta}_t$ in the regression model is estimated separately for any 34 estimation areas and dwelling-type for period *t*. Parameters $\beta_{01t}, \dots \beta_{0k_1t}$ represent stratum effects in Helsinki in period *t*. The k-dimensional vector $\boldsymbol{x}'_{it}$ consists of exogenous independent variables (i.e. quality characteristics). Term ε_{it} is a random error term, which does not contain systematic information about the data generating process of prices. It is assumed, that $E\left(\varepsilon_{it} \middle| \boldsymbol{x}'_{it}\right) = 0$ and $Var\left(\varepsilon_{it} \middle| \boldsymbol{x}'_{it}\right) = \sigma_t^2 < \infty$. In our model specification, the error covariance matrix is diagonal – the most natural situation for heterogeneously behaving cross-sectional data (i.e., 34 estimation areas in time).

The estimation of unknown parameters follows the ordinary-least-squares (OLS) method. The OLS estimators are obtained by minimizing the residual sum of squares using basic principles of Gauss LS-regression. We do the estimation using a twostep OLS method (see Davidson & MacKinnon, 1993, p. 19-25), where we transform observations into deviation of means with respect to our partition. In step one, we get estimates for our unknown parameters that we denote as $\widehat{\boldsymbol{\beta}}_t$. In the second step partition-specific or stratum effects $\hat{\beta}_{0kt}$, for stratum *k* are estimated as

$$(2) \qquad \hat{\beta}_{0kt} = log(\bar{p}_{kt}) - \bar{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_t, \text{ for } k = 1,\dots, k_1.$$

where $\bar{p}_{kt} = \prod p_{ikt}^{1/n}$ present unweighted geometric average price for stratum *k* and $\bar{\boldsymbol{x}}'_{kt}$-vector unweighted arithmetic averages of quality characteristic (see Koev, 2003, p. 22-26). According to the Frisch, Waugh and Lovell -theorem (Davidson & MacKinnon, 1993), the OLS –estimation of the slopes can always be carried out via centralized variables. The constant term for stratum *k* is estimated by forcing the regression plane through the point of averages (algebraic property of Gauss regression). This method is computationally extremely effective especially when partition includes hundreds/thousands of strata (see Suoperä & Vartia, 2011). The semilogarithmic equation (1) estimated by the OLS takes the form (here the first equation corresponds to Helsinki region)

$$(3) \qquad \log(p_{it}) = \hat{\beta}_{01t} + \cdots + \hat{\beta}_{0k_1t} + \boldsymbol{x}'_{it}\widehat{\boldsymbol{\beta}}_t + \hat{\varepsilon}_{it}.$$

The OLS estimators and the equation (3) - simply log-prices, quality characteristics and their arithmetic and geometric averages - is everything that is needed to construction a hedonic index numbers for any stratum, here for strata in Helsinki.

The estimation of equation (1) is based on equal weights for all observations. The method is traditionally used and is clearly interpretable. Koev (2003) underlines that the statistical properties of price changes based on unweighted geometric average prices and their Oaxaca decompositions is also clearly interpretable. This means that quality adjusting combined with index numbers have been derived by algebra and therefore are based on easily understandable transparent mathematics. This is a fine property of hedonic methods introduced by Koev (2003).

Another principle using equations (1) – (3) is to weight observations in equation (3) by '*a weighted-by-economic-importance*'-variable and aggregate price model into stratum-aggregates. For examples of this method see Suoperä (2004, 2006); Suoperä & Auno (2021); Suoperä, Nieminen, Montonen and Markkanen (2021). This approach satisfies also basic algebraic properties of Gauss LS-regression and leads to identical Oaxaca decompositions as in Koev (2003), but instead of unweighted geometric means they are based on weighted arithmetic average prices. The method used in these studies is a standard practice in survey studies – first estimate and then weight. Applications of Koev and Suoperä are seminal hedonic methods combining regressions and index numbers. The methods are presented in Chapter 4.

3.3 The Aggregation Stage

In regression analysis stage transacted dwellings are 'split' locally and by apartment-type into two nested partitions. The stratums in both partitions are grouped regionally into 34 separate estimation equations. These equations – price models – are specified as semilogarithmic having flexible functional form. Table 1 belove describes explanatory variables used in regional price models. Price modelling for dwellings is specified to be heterogeneously behaving cross-sections which means estimation of thousands of unknown parameters. How to summarize this 'huge mass of statistical information'? For that we use the method that was introduced in Suoperä and Vartia (2011). We use the following steps: 34 separately estimated equations (3) for terraced houses and blocks of flats include thousands of different behaviors for every year. They are summed up, and after that making a solution backward into observation level, we get first the representative aggregate equation that is common/representative for all behaviors and second individual deviations of it as heterogeneity behaviors. To obtain the standard errors, we estimate the model again with OLS. We do this because obtaining the standard errors would be difficult otherwise (Suoperä and Vartia, 2011, p. 11-18).

Table 3.1: The exogenous variables used in the regional price models for terraced houses and block of flats in Finland.

Variable	Description
Dwelling type dummies	Classify observations into four dwelling types: one-room, two-rooms, three-rooms or more and terraced houses for each region
x_1	Square meters of the transacted dwelling
$x_2 = sqrt(x_1)$	Square root of the square meters of the transacted dwelling
x_3	Age of dwelling in years is calculated here as deviation from the year 2020 (as in Koev, 2003, p.39), but will be changed every year when the base year is changed.
$x_4 = sqrt(x_3)$	Square root of age
x_5	Driving time by car to the nearest local centre. Centres defined by the Finnish Environment Institute are areas that have dense and versatile services, such as shops, leisure services, public services, as well as jobs in various industries and habitation.
$x_6 = sqrt(x_5)$	Square root of the average driving time by car to the nearest local centre.
x_7	Owner of the building lot: Dummy variable that gets value 0, when the building lot is rented and otherwise 1.

Before we present empirical results of the synthesis stage, we test the significance of the additional partition based on postal areas. Practically this means that we must make statistical inference of a set of 34 equations according two partitions. This is done separately for terraced houses and blocks of flats. Our hypothesis is clear: Additional stratification by postal areas is unnecessary meaning restricting them zero in estimation. This leads to two sum of squared errors – one for set of free models and another to restricted ones. The most natural test for that is familiar F-test, that is

$$F\sim\ \{(SSE_0-SSE_1)/R\}/\{SSE_1/(N-K)\}$$

where SSE_0 is the sum of squared errors of the restricted model, SSE_1 is the sum of squared errors of the free model, $(N-K)$ is the degrees of freedom of the free model and R is the number of linear restrictions.

When the degrees of freedom for free model becomes large - here for terraced houses and block of flats more than 15000 and 35000 respectively – the F-statistics reduced into χ_R^2-test, where R corresponds number of linear restrictions (see Greene 1997, p. 344 and p. 657). For example, a 1% critical value of $\chi_{60}^2=1.46$ and becomes closer to one when R > 60. Table 3.2 shows results for testing the significance of additional partition based on postal areas.

Table 3.2: The values of the F–test statistics in testing the hypothesis of the homogeneity of partitions. The number of linear restrictions, R, are in parenthesis.

Year	2015	2016	2017	2018	2019	2020
Terraced houses	11.9	11.6	10.5	10.3	10.0	10.3
Block of flats	11.7	12.2	10.4	11.2	11.6	12.4

The above Table tells that the hypothesis is always highly significant and rejected. If the additional partition is omitted from price modelling, the price models necessarily lead into biased estimation of behavioral beta-parameters because of omission of important variables - here detailed postal area partition.

Now we have knowledge of statistical inference of our price models. In the following two tables we show empirical results based on the analysis and synthesis technique similarly as in Suoperä and Vartia (2011).

The estimation results for terraced houses are convincing. Two last estimates and their standard errors (i.e., $He(\alpha)$ and $He(x\beta)$) tells the same story as F-test statistics – additional partition based on postal areas is extremely significant for all estimations. Similar results hold also for behavioral heterogeneity $He(x\beta)$.

The estimation results for block of flats are even more convincing compared to terraced houses. The same holds also for heterogeneity behaviors, $He(\alpha)$ and $He(x\beta)$, which are extremely significant for all estimations.

Using analysis and synthesis stages for estimations of heterogeneously behaving cross-sections similarly as in Suoperä and Vartia (2011) more than 10000 parameters and their standard errors and other statistics may be presented simply by two tables. Tables will adequately tell how reliable our price modelling really are.

Table 3.3: Estimation results for the price equations for terraced houses using analysis and synthesis-technique developed in Suoperä and Vartia (2011).

Year	2015	2016	2017	2018	2019	2020
Obs	17048	16700	17133	16930	17248	19079
Stratum	1348	1351	1363	1342	1357	1357
Adjr2	0.865498	0.86624	0.859484	0.852368	0.847891	0.83561
RMSE	0.166412	0.16755	0.177312	0.192414	0.20241	0.220914
Constant	8.717586	8.668361	8.659984	8.682045	8.522123	8.744483
se(c)	0.037465	0.038702	0.039722	0.043469	0.044198	0.045708
x_1	-0.00115	-0.00064	-0.00263	-0.002	-0.00208	-0.0006*
se(x_1)	(0.000451)	(0.000464)	(0.000477)	(0.000521)	(0.000531)	(0.000551)
x_2	-0.0276	-0.03508	-0.00695	-0.00822	-0.00558	-0.03966
se(x_2)	(0.008051)	(0.008328)	(0.008587)	(0.009398)	(0.0096)	(0.010009)
x_3	0.008345	0.006184	0.007471	0.008012	0.004229	0.006676
se(x_3)	(0.000372)	(0.000396)	(0.000372)	(0.000385)	(0.000411)	(0.000382)
x_4	-0.20295	-0.17823	-0.19322	-0.20057	-0.16173	-0.18276
se(x_4)	(0.004288)	(0.00444)	(0.004175)	(0.004259)	(0.004353)	(0.004014)
x_5	-0.00737	-0.0078	-0.00316	0.002768	-0.00069	0.00107
se(x_5)	(0.000536)	(0.000494)	(0.00047)	(0.000384)	(0.000615)	(0.000805)
x_6	0.005297	0.009539	-0.01416	-0.05397	-0.03171	-0.03407
se(x_6)	(0.003965)	(0.003704)	(0.003597)	(0.003222)	(0.004547)	(0.00568)
x_7	0.04604	0.050197	0.039587	0.053638	0.078146	0.073761
se(x_7)	(0.003322)	(0.003354)	(0.00351)	(0.003778)	(0.004008)	(0.004131)
$He(\alpha)$	1	1	1	1	1	1
se($He(\alpha)$)	(0.003756)	(0.003593)	(0.003681)	(0.003788)	(0.003871)	(0.003963)
$He(x\beta)$	1	1	1	1	1	1
se($He(x\beta)$)	(0.004644)	(0.00419)	(0.004552)	(0.004622)	(0.004385)	(0.00525)

Note: all parameters are statistically significant with 99% confidence, with the exception of the parameters highlighted in yellow.

Table 3.4: Estimation results for the price equations for block of flats using analysis and synthesis-technique developed in Suoperä and Vartia (2011).

Year	2015	2016	2017	2018	2019	2020
Obs	38219	38119	38285	37052	37115	38813
Stratum	1733	1723	1721	1710	1711	1723
Adjr2	0.905216	0.905714	0.902352	0.897711	0.899173	0.899524
RMSE	0.185637	0.190408	0.1966	0.209496	0.215889	0.223601
Constant	10.69365	10.72264	10.72156	10.69992	10.63826	10.4095
se(c)	0.024008	0.024911	0.024742	0.026185	0.026652	0.026329
x_1	0.017103	0.018156	0.018412	0.017935	0.019187	0.017345
se(x_1)	(0.000372)	(0.00039)	(0.000386)	(0.000412)	(0.000425)	(0.000436)
x_2	-0.34458	-0.36064	-0.36816	-0.35858	-0.37508	-0.34277
se(x_2)	(0.005675)	(0.005993)	(0.005973)	(0.006379)	(0.006594)	(0.006727)
x_3	0.020625	0.020378	0.018763	0.019579	0.017837	0.015518
se(x_3)	(0.000257)	(0.000252)	(0.000253)	(0.000259)	(0.000251)	(0.000235)
x_4	-0.34753	-0.34436	-0.32478	-0.33085	-0.30482	-0.26915
se(x_4)	(0.003273)	(0.003176)	(0.003129)	(0.003167)	(0.003005)	(0.002726)

x_5 se(x_5)	-0.00015 (0.000708)	-0.00383 (0.000723)	-0.00625 (0.0008)	-0.00466 (0.000875)	-0.01101 (0.000845)	-0.00797 (0.000847)
x_6 se(x_6)	-0.06122 (0.003844)	-0.04783 (0.003901)	-0.03421 (0.00425)	-0.04147 (0.004615)	-0.01246 (0.004499)	-0.03729 (0.00454)
x_7 se(x_7)	0.067217 (0.002501)	0.067989 (0.002501)	0.06661 (0.002645)	0.060394 (0.002811)	0.077362 (0.002877)	0.087561 (0.002915)
$He(\alpha)$ se($He(\alpha)$)	1 (0.002845)	1 (0.002587)	1 (0.002994)	1 (0.002717)	1 (0.002683)	1 (0.002777)
$He(x\beta)$ se($He(x\beta)$)	1 (0.002098)	1 (0.002041)	1 (0.002086)	1 (0.002153)	1 (0.002122)	1 (0.002057)

Note: all parameters are statistically significant with 99% confidence, with the exception of the parameters highlighted in yellow.

The equation (1) has non-linear square root terms in number of square meters, age and distance in driving times. To interpret the estimation results of eq. (3) we take partial derivatives with respect to number of square meters (x_1), age (x_3) and distance (x_5) of transacted dwelling, that is (here postal area *r*)

$$\frac{\partial \log(p_{irt})}{\partial x_{i1rt}} = \hat{\beta}_{1rt} + \hat{\beta}_{2rt}/sqrt(x_{i1rt}), \forall\, i \in A_r,$$

$$\frac{\partial \log(p_{irt})}{\partial x_{i3rt}} = \hat{\beta}_{3rt} + \hat{\beta}_{4rt}/sqrt(x_{i3rt}), \forall\, i \in A_r \text{ and}$$

$$\frac{\partial \log(p_{irt})}{\partial x_{i5rt}} = \hat{\beta}_{5rt} + \hat{\beta}_{6rt}/sqrt(x_{i5rt}), \forall\, i \in A_r.$$

When we calculate cumulative sums of partial derivates for ordered cohorts (i.e. x_{i1}, x_{i3} are ordered starting from smallest) and get Figures 1 to 3.

Figure 1: The price effect of size (log-%) on the square meter price of apartment (year 2020).

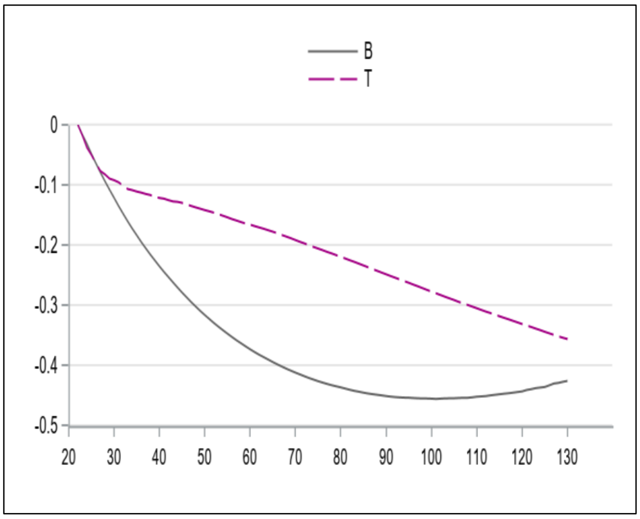


Figure 2: The price effect of age (log-%) on the square meter price of apartment (year 2020).

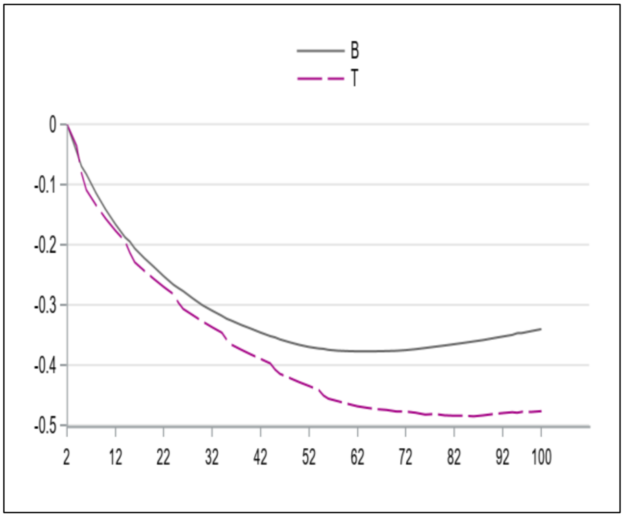
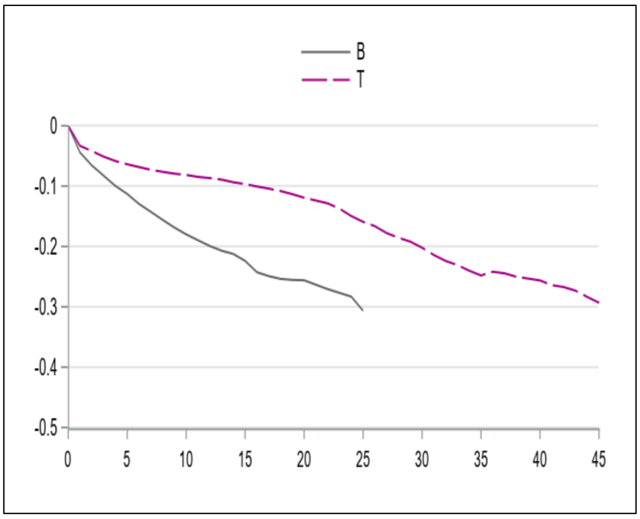


Figure 3: The price effect of distance (log-%) on the square meter price of apartment (year 2020). Driving time to the nearest local centre in minutes.



In Figure 1 we see that square meter prices decline differently for terraced houses (T) and block of flats (B) when number of squares increase. For block of flats square meter price of a 60 m^2 dwelling is about 30 – 35 log-% lower compared to a 20 m^2 flat - for terraced houses only about 15 log-%. Increase of age will decline prices quite similarly for terraced houses and block of flats up to 40 - 50 years old dwellings, but after that differently – increase of age of old block of flats in town centers increases prices slightly, but not for terraced houses. The Figure 3 tell that increase of driving time from center decline prices differently for these dwelling types. The same 30 log-% decline of prices is achieved for block of flats by 25 minutes and for terraced houses by 45 minutes. This is caused different location of dwelling types – block of flats nearby center and terraced houses mostly in fringes of centers.

The estimation results are summarized as:

1. Almost all parameters for representative behavior are statistically significant (statistically insignificant parameters are highlighed with yellow color at 99 % level).
2. Postal area indicators ($He(\alpha)$) are strongly significant for terraced houses and block of flats and must be included into the price models. Also, separate estimation for both dwelling type is necessary.
3. The data should be analyzed by heterogeneously behaving cross-section (i.e. $He(x\beta)$).
4. All quality characteristics (i.e., x_1, x_3, x_5) have negative effect on prices (see Figures 1-3). The owner of the building lot-dummy variable have positive and significant effect on square meter prices for both dwelling types.

The analysis and synthesis stages include all that is necessary for construction of hedonic index numbers. We need estimated price equations at observation level, knowledge about basic algebraic properties of estimation method, knowledge about consistent and unbiased aggregation rules including calibration of observations, knowledge of conditional and unconditional averages and knowledge of index number formulas. Next, we show how these ‘parts of whole’ are combined together by simple and transparent algebra.

4 Combining Regression Analysis and Index Numbers

When bilateral or multilateral methods are not available for commodities comparable quality, a well-known time-dummy hedonic regression is often used to resolve the problem of quality adjusting. However, as mentioned, there are issues with its use. Namely, the statistical properties of the time-dummy estimator maybe ambiguous, when the regression includes weights by economic importance (see Summers (1973); Rao (2004) and Diewert and Fox (2018)). Moreover, the link with the traditional index number theory is somewhat unclear, i.e. how to proceed with aggregation. We propose a solution for quality adjusting based on simple statistics, some algebra and hedonic index numbers, which in our view is preferable for statistical offices, since it is transparent, minimizes modeling assumptions, and is consistent with index number tradition of consistency in aggregation. Our analysis herein follows the tradition of Koev (2003); Suoperä (2004, 2006); Vartia, Suoperä & Vuorio (2021) and Suoperä & Auno (2021).

In the first step, we have two partitions for price models and two consistent aggregation rules for them. We take logical steps, starting from aggregation of observations of equation (3) into strata for two partitions using two aggregation rules – first being unweighted and second *weighted-by-economic-importance*-variable. We provide an example using one arbitrarily selected stratum and time periods (0, *t*), but similar process can be followed for all other strata.

The second step is the well-known decomposition introduced by Oaxaca (1973) for stratum aggregates – here for two partitions and two aggregates (unweighted and weighted). Two partitions are necessary – we show that inadequate stratification of transacted dwellings leads to biased change estimates including not only true price changes but also quality changes that are not controlled by the hedonic model well enough. This is done using additive decomposition of value change.

The last step is similar as traditional index numbers – the averaged stratum-level price decompositions are summed up using weights of index number formulas, that is '*weights-by-economic-importance*'-variable. We analyze two sets of index number formulae. The first set is based on formulas using old or new weights (asymmetrical weights) and are called as a basic set of index numbers. Laspeyres (*L*) and Log-Laspeyres (*l*) uses base period weights (i.e. old weights) and Log-Paasche (*p*) and Paasche (*P*) instead uses observation period weights (i.e. new weights). The second set of index numbers include four formulae using symmetrical weights: Montgomery-Vartia (*MV*), Törnqvist (*T*), Fisher (*F*) and Sato-Vartia (*SV*). We call these index number formulae as *excellent*. For the fundamental analysis of these index number formulae see Vartia & Suoperä, 2018. The analysis therein is in logarithmic form.

4.1 Aggregation of Observations into Strata

We analyze two aggregation rules for observations. The first is traditional unweighted geometric average, where a *weighted-by-economic-importance*-variable is neglected. This means that all transacted dwellings get an equal weight in aggregation. The second aggregation rule is weighted arithmetic average, say unit price, that has been derived in this study to be logarithmic representation of unit value similarly as in Suoperä (2004, 2006 p.2-5, Annex 5); Vartia, Suoperä & Vuorio (2021) and Suoperä & Auno (2021, p. 3-7). Fundamental analysis is shown in these papers and is based on utilization of *logarithmic mean* developed by Törnqvist (L. Törnqvist, 1935, p. 35; Y. Vartia, 1976; L. Törnqvist, P. Vartia & Y. Vartia, 1985, p. 44).

In the previous chapter we specify and estimate 34 price models separately for terraced houses and block of flats. For notational simplicity we take only one equation (no subindex for equation *j* = 1,…,34) and explain the most important algebraic properties of the equation (3), when the equation is estimated using a standard OLS-method with equal weights (Greene, 1997, p. 243-244):

1. Aggregation of observations for stratum *k* leads to conditional average

   $log(\bar{p}_{kt}) = \hat{\beta}_{0kt} + \bar{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_t$, for $k = 1,\ldots,k_1$, where $\bar{p}_{kt} = \prod p_{ikt}^{1/n_k}, i = 1, \ldots, n_k$,

   meaning that regression hyperplane passes through the means of dependent and independent variables.

2. Residuals sum up to zero for all $k = 1,\ldots,k_1$ (unbiased estimators).

3. The average of fitted values of prices (conditional average) equals the average of actual prices (unconditional average) in all stratums k = 1,…, k_1.

The properties of the OLS are based on equal weighting of observations. Koev (2003) uses in his study these three properties and derive a hedonic quality adjusting method that is based on index numbers. These three properties combined with the best linear unbiased estimator, BLUE (homoscedastic errors), forms the basis for hedonic quality adjusting that is hard to beat. These properties are derived using consistent aggregation rules (CA) of our averages, unbiased statistics and index numbers with symmetric or asymmetric weights.

We see the WTPD-model (Summers 1973; Diewert and Fox, 2018) – estimate and weight simultaneously - approach problematic because of asymmetric weighting variable, for example value- or quantity shares, depending on unit values (i.e., unit prices as dependent variable); first causing unknown properties (bias) of beta estimates and that's why second unknown bias for quality adjusting. Our method is based on a very simple 'Estimate first and then weight' approach.

We analyze another stratum aggregate for equation (3) – weighted arithmetic average that satisfies above three algebraic properties of the OLS method. The derivation follows Suoperä (2004, 2006), Vartia, Suoperä and Vuorio (2021), and Suoperä and Auno (2021) and is based on a simple idea used in survey studies: *First estimate equations (3) separately and then weight them in aggregation.* In derivation, we need values, quantities and prices (i.e., v_{ikt} , q_{ikt} and p_{ikt}) and a theorem of logarithmic mean. Logarithmic mean is defined for two positive figures *x* and *y* as follows (L. Törnqvist, 1935, p. 35; Y. Vartia, 1976; L. Törnqvist, P. Vartia and Y. Vartia, 1985, p. 44)

$$L(x,y) = (x-y)/\log(x/y)\,, if\ x \neq y \text{ and}$$

$$= x,\ if\ x = y$$

Another useful representation is $\log(y/x) = (y\text{-}x)/L(x,y)$ meaning that the log change from *x* to *y* is a relative change compared to the logarithmic mean. This indicator of relative change is a ratio that is symmetrical, additive and independent of measurement unit and that is why it may applied for sets of positive values of *x* and *y*. Let us now analyze values, quantities and prices for dwellings located in arbitrary stratum A_k in time period *t*, that is $\{v_{ikt}, q_{ikt}, p_{ikt}\}$ and define their logarithmic mean as follows

$$L(\textstyle\sum_i v_{ikt}, \sum_i q_{ikt}) = \sum_i \frac{v_{ikt}-q_{ikt}}{\log(\sum_i v_{ikt}/\sum_i q_{ikt})}, \text{ or } \log(\sum_i v_{ikt}/\sum_i q_{ikt}) = \sum_i \frac{v_{ikt}-q_{ikt}}{L(\sum_i v_{ikt}, \sum_i q_{ikt})}.$$

Using above equations and some algebra we get (use definition of weighted arithmetic average $\sum_i v_{ikt}/\sum_i q_{ikt} = \bar{p}_{kt}$ and definition of unit value $v_{ikt}/q_{ikt} = p_{ikt}$, see also definition of logarithmic mean)

$$\text{(4)} \qquad \log(\textstyle\sum_i v_{ikt}/\sum_i q_{ikt}) = \sum_i \frac{L(v_{ikt}, q_{ikt})}{L(\sum_i v_{ikt}, \sum_i q_{ikt})} \log(v_{ikt}/q_{ikt}) \leftrightarrow$$
$$\log(\bar{p}_{kt}) = \textstyle\sum_i \frac{L(v_{ikt}, q_{ikt})}{L(\sum_i v_{ikt}, \sum_i q_{ikt})} \log(p_{ikt})$$

The above equation is a logarithmic representation of unconditional weighted arithmetic average. The corresponding conditional average is derived in Suoperä (2006, p 4-5, Annex 5). All these statistics are unbiased and consistent in aggregation (*CA*). These statistics satisfy properties 1 to 3. For unweighted geometric average, unconditional and conditional averages are trivially equal. Using '*a weighted-by-economic-importance*'-variable w_{ikt} in (4) to derive estimates for weighted arithmetic average leads to reparameterization of the price model. For this reparametrized model $\bar{p}_{kt} = \prod p_{ikt}^{w_{ikt}} = \bar{\bar{p}}_{kt}$ holds exactly even the weights are not independent of measurement units. In Table 4.1 we collect together the most important statistics for which we apply hedonic quality adjusting and index numbers.

Table 4.1: Important statistics for hedonic quality adjusting.

Average	Unconditional	Conditional
Unweighted geometric average	$\bar{p}_{kt} = \prod p_{ikt}^{w_{ikt}}$, where $w_{ikt} = 1/n_k$, for all $i \in A_k$	$log(\bar{p}_{kt}) = \hat{\beta}_{0kt} + \bar{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_t$ or by portioned vectors $log(\bar{p}_{kt}) = (1{:}\,\bar{\boldsymbol{x}}_{kt})'\begin{pmatrix}\hat{\beta}_{0kt}\\ \widehat{\boldsymbol{\beta}}_t\end{pmatrix} = \bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_t^*.$
Weighted arithmetic average	$\bar{\bar{p}}_{kt} = \prod p_{ikt}^{w_{ikt}}$, where $w_{ikt} = \frac{L(v_{ikt}\,,q_{ikt}\,)}{L(\sum_i v_{ikt},\sum_i q_{ikt})}$, for all $i \in A_k$	$\log(\bar{\bar{p}}_{kt}) = \check{\beta}_{0kt} + \bar{\bar{\boldsymbol{x}}}'_{kt}\widehat{\boldsymbol{\beta}}_t$ or by portioned vectors $log(\bar{\bar{p}}_{kt}) =(1{:}\,\bar{\bar{\boldsymbol{x}}}_{kt})'\begin{pmatrix}\check{\beta}_{0kt}\\ \widehat{\boldsymbol{\beta}}_t\end{pmatrix} = \bar{\boldsymbol{z}}'_{kt}\check{\boldsymbol{\beta}}_t^*,$ where $\check{\beta}_{0kt} = \log(\bar{\bar{p}}_{kt}) - \bar{\bar{\boldsymbol{x}}}'_{kt}\widehat{\boldsymbol{\beta}}_t,$ where $\bar{\bar{\boldsymbol{x}}}'_{kt} = \sum_{\boldsymbol{i}} w_{ikt}\,\boldsymbol{x}'_{ikt}$

4.2 Algebra of Price-Ratio Decompositions

In the previous chapter unbiased estimates take an important role: They are unbiased estimates of unknown parameters and unbiased unconditional and conditional averages – here unweighted geometric and weighted arithmetic averages. Now we show how these statistics may be used in hedonic quality adjusting applying them to a well-known decomposition developed by Oaxaca (1973). The decomposition is not unique but can be constructed consistently and similarly as in Koev (2003) and Suoperä (2004, 2006).

Following the example in Koev (2003), we take two time periods, the base period (*t* = 0, a previous year) and the observation quarter of current year (*t*) and only one stratum A_k. We use vector notations for our conditional and unconditional average prices and calculate the difference between two price models (0, *t*) in spirit of Oaxaca. The algebra is presented only for unweighted geometric average and may be deduced analogously for weighted arithmetic average. The differences in prices are given as

(5a) $$log(\bar{p}_{kt}) - log(\bar{p}_{ko}) = \bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_t^* - \bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_0^* \leftrightarrow$$

(5b) $$log(\bar{p}_{kt}/\bar{p}_{k0}) = \left(\bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_0^* - \bar{\boldsymbol{z}}'_{k0}\widehat{\boldsymbol{\beta}}_0^*\right) + \left(\bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_t^* - \bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_0^*\right).$$

We are playing with average prices and some interpretations are needed. First, $\bar{\boldsymbol{z}}'_{k0}\widehat{\boldsymbol{\beta}}_0^* = log(\bar{p}_{ko})$ and $\bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_t^* = log(\bar{p}_{kt})$, where $\bar{p}_{kt} = \prod p_{ikt}^{w_{ikt}}$, where $w_{ikt} = 1/n_k$, for all $i \in A_k$ (see Table 4.1). The term $\bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_0^* = log(\tilde{p}_{kt})$ is an estimated/imputed average price of observation period quality variables (i.e., $\bar{\boldsymbol{z}}'_{kt}$) using base periods valuation of characteristics (i.e., $\widehat{\boldsymbol{\beta}}_0^*$). Collecting these terms together we get

(6) $$log(\bar{p}_{kt}/\bar{p}_{k0}) = log(\tilde{p}_{kt}/\bar{p}_{k0}) + log(\bar{p}_{kt}/\tilde{p}_{kt})$$

The equation (6) is very simple and holds as an identity. On the left, we have the price-ratio of actual average prices (here conditional being equal with unconditional average), and on the right, the first term is quality correction estimated using base period valuation of characteristics (i.e., $log(\tilde{p}_{kt}/\bar{p}_{k0}) = (\bar{\boldsymbol{z}}'_{kt} - \bar{\boldsymbol{z}}'_{k0})\widehat{\boldsymbol{\beta}}_0^*$) and the second is quality adjusted price change (i.e, $log(\bar{p}_{kt}/\tilde{p}_{kt}) = \bar{\boldsymbol{z}}'_{kt}\left(\widehat{\boldsymbol{\beta}}_t^* - \widehat{\boldsymbol{\beta}}_0^*\right)$) estimated with comparable in quality (i.e., $\bar{\boldsymbol{z}}'_{kt}$, for all *k* and *t*).

The equations (5a) and (5b) includes the insight in Koev (2003): First, estimation of price models (here 34 equations for both dwelling-types) should estimate only for base period - $\bar{\boldsymbol{z}}'_{kt}\widehat{\boldsymbol{\beta}}_t^* = log(\bar{p}_{kt})$ holds trivially meaning that conditional and unconditional averages are equal for any *t* and *k* (i.e., algebraic property of the OLS-method). Second, quality adjusting is based on simple and transparent algebra. Third, the quality corrections may carry out separately for any characteristics of the price model, that is

$$log(\tilde{p}_{kt}/\bar{p}_{k0}) = (\bar{\boldsymbol{z}}'_{kt} - \bar{\boldsymbol{z}}'_{k0})\hat{\boldsymbol{\beta}}_0^* = (1:\bar{\boldsymbol{x}}_{kt})'\begin{pmatrix}\hat{\beta}_{0kt}\\ \hat{\boldsymbol{\beta}}_0\end{pmatrix} - (1:\bar{\boldsymbol{x}}_{k0})'\begin{pmatrix}\hat{\beta}_{0kt}\\ \hat{\boldsymbol{\beta}}_0\end{pmatrix}$$

$$= (\bar{x}_{1kt} - \bar{x}_{1k0})\hat{\beta}_{10} + \cdots + (\bar{x}_{Rkt} - \bar{x}_{Rk0})\hat{\beta}_{r0},$$

where subscript r =1,…,R refers exogenous variables of the price model. The equation above tells us that when $\bar{x}_{rkt} - \bar{x}_{rk0} = 0$, for all ($r$, t, 0), the quality adjusting is unnecessary, and equation (6) reduces to a very simple logarithmic price-ratio of average prices – in terminology of traditional index calculation into classification index. When $\bar{x}_{rkt} - \bar{x}_{rk0} \neq 0$, for all ($r$, t, 0) quality correction may perform for single variable alone or some reasonable combinations quality characteristics (here for average age, average of square meters, average driving time and average of owner of building lot).

The equation (6) is in center for which index number formulas are applied. We analyze formulae that have asymmetric and symmetric weights and apply them for two partitions and two price-concept - unweighted geometric and weighted arithmetic averages. The analysis is done using logarithmic representations of formulae and price-decompositions.

4.3 Index Number Formulas

Index number theory begins by aggregating the decomposition (6). We adopt two sets of index number formulae. The first set is based on formulae using asymmetric old (Laspeyres *La* and Log-Laspeyres *l*) or new weights (Log-Paasche *p* and Paasche *P*). The second set of index numbers includes four formulae: Montgomery-Vartia (*MV*), Törnqvist (*T*), Fisher (*F*) and Sato-Vartia (*SV*). These are based on symmetrical weighting and are called as *excellent formulae* (see Vartia & Suoperä, 2017, 2018).

In Table 4.2 we gather all information that is necessary for calculation of hedonic price indices. All index number formulae are represented in logarithmic form, including Laspeyres, Paasche and Fisher (see Vartia, 1976, p.128). Practically this means, that the aggregation of price changes or its decomposition in (6) is always done much simpler in additive form and then transformed back to indices.

Table 4.2: Weights for index number formulae (logarithmic forms).

Basic formulae, see Vartia & Suoperä, 2017, 2018, L means logarithmic mean, see Vartia, 1976a, p. 128	
Symbol and name of formula	Weights of the formula
Laspeyres, f = L	$w_{k,f} = w_{k,L}^0 = L(p^t q^0, p^0 q^0)$
log-Laspeyres, f = l	$w_{k,f} = w_{k,l}^0 = v_k^0/V^0$
log-Paasche, f = p	$w_{k,f} = w_{k,p}^t = v_k^t/V^t$
Paasche, f = P	$w_{k,f} = w_{k,P}^t = L(p^t q^t, p^0 q^t)$
Excellent formula, see Vartia & Suoperä, 2017, 2018), L means logarithmic mean, see Vartia, 1976	
Törnqvist, f = T	$w_{k,f} = \bar{w}_{k,T} = 0.5 \cdot (w_k^0 + w_k^t)$
Sato-Vartia, f = SV	$w_{k,f} = \bar{w}_{k,SV} = \dfrac{L(w_k^t, w_k^0)}{\sum L(w_k^t, w_k^0)}$
Montgomery-Vartia, f = MV	$w_{k,f} = \bar{w}_{k,MV} = L(p^t q^t, p^0 q^0)$
Fisher, f = F	$w_{k,f} = \bar{w}_{k,F} = 0.5 \cdot (L(p^t q^0, p^0 q^0) + L(p^t q^t, p^0 q^t))$

Applying weights for equation (6) we get logarithmic representations for formulae, that is

$$log\left(P_{f,A}^{t/0}\right) = \sum\nolimits_{k} w_{k,f}\, log(\bar{p}_{kt}/\bar{p}_{k0})$$

$$= \sum\nolimits_{k} w_{k,f}\, log(\tilde{p}_{kt}/\bar{p}_{k0}) + \sum\nolimits_{k} w_{k,f}\, log(\bar{p}_{kt}/\tilde{p}_{kt})$$

or

(7a) $log\left(P_{f,A}^{t/0}\right) = log\left(P_{f,QC}^{t/0}\right) + log\left(P_{f,QA}^{t/0}\right)$ and index numbers

(7b) $P_{f,A}^{t/0} = P_{f,QC}^{t/0} \cdot P_{f,QA}^{t/0}$,

where subscript *f* notes formula, *A* actual price change of averages, *QC* quality corrections and *QA* quality adjusted price change. If one likes, quality corrections can be decomposed *variable-by-variable* such that $P_{f,QC}^{t/0} = P_{f,QC,x_1}^{t/0} \cdot P_{f,QC,x_2}^{t/0} \cdot \ldots \cdot P_{f,QC,x_R}^{t/0}$, which holds as an identity. Weighting means here always '*a weighted-by-economic-importance*'-variable familiar to index numbers.

5 Empirical Results

Following table presents the focus of our study. Some words are necessary. In Chapter 3 we define two nested partitions and price models for them. *In the first step*, estimated price models are aggregated from observation level into four classes of Table 4.3.

Table 4.3: Focus of the study.

Partition\Aggregation rule	Unweighted geometric average	Weighted arithmetic average
Partition one	1	2
Partition two	3	4

In the second step, we estimate the price decomposition (6) for classes 1 to 4. To make comparisons possible between two partitions, we must aggregate class 3 and 4 into level of partition one. The aggregation is done using equation (7a) and index number formulae in Table 4.2 and Jevons (i.e., equal weights). This step gives information about differences between formulae and necessity of detailed partition two. Also, when the index number formula *f* decomposes the value change into price and quantity changes (i.e. $log\left(V^{t/0}\right) = log\left(P^{t/0}\right) + log\left(Q^{t/0}\right)$, i.e., additive decomposition of value change, $ADVC$) for all aggregation levels, we get a simple explanation to the difference between the change of unit value in class 2 compared to hedonic price index *f* for class 4 aggregated into partition one (see Suoperä & Auno, 2021, p. 11). In the third step we ask, 'How closely class 3 and 4 are related?' We simply regress price changes (or index series) for strata in partition two in class 4 on corresponding price changes in class 3. This simple regression tells correlation between class 3 and 4 and empirical analysis ends here.

All index numbers and index series are based on base strategy, where the base period is a previous year normalized as an average quarter and the observation period is a quarter of a current year.

5.1 Does the Formula Matter?

Population quantities are estimated for partition one, and that is where we have to explore whether the formula matters. We compare classes (1, 3) and (2, 4) from table 4.3. First for pair (1, 3): In class 1 aggregation is done directly from observations into stratums and in class 3 first into partition 2 and then from partition 2 into strata of partition 1 using the index number formulae. The same method is applied for pair (2, 4). Statistical inference of price models strongly indicates to use partition 2 – since the explanatory power increases and quality adjustment improves when more categories are included in the model. Let us now use the theory of index numbers for answering this question. The results are presented in following figures separately for terraced houses and blocks of flats.

5.1.1 Synthesis of Unweighted Geometric Average, UWGA

Figure 1: Hedonic index series for actual average prices in stratum ‘Espoo, subarea 1, three-rooms’. Basic index Numbers from 2015.0 to 2020.4 (P(2015)=1). Terraced houses.

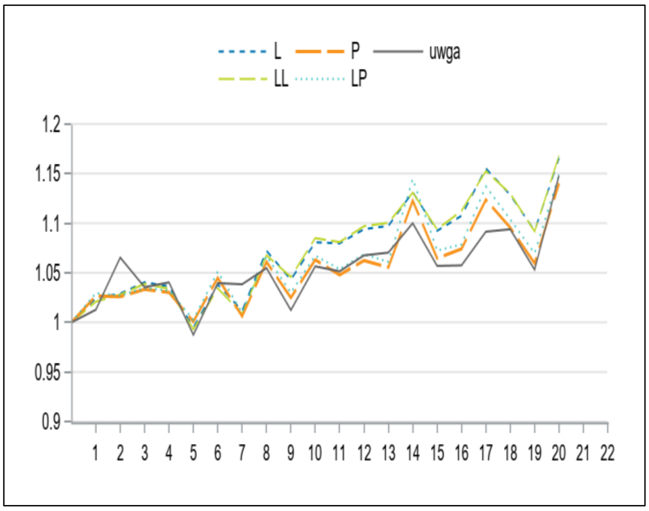


Figure 2: Hedonic index series for actual average prices in stratum ‘Espoo, subarea 1, three-rooms’. Excellent index numbers from 2015.0 to 2020.4 (P(2015)=1). Terraced houses.

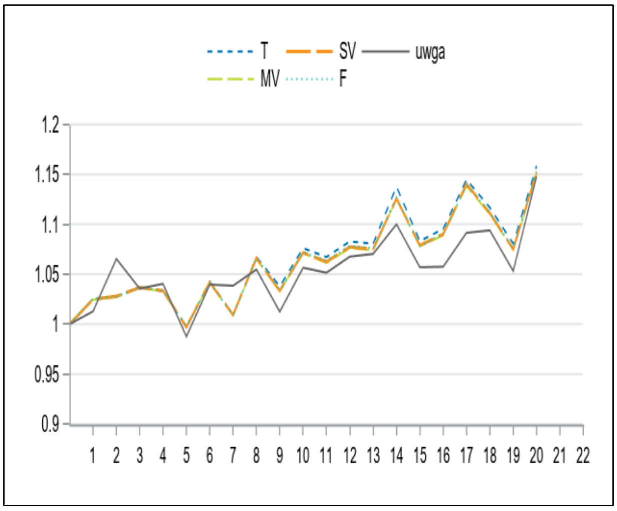


Figure 3: Hedonic index series for actual average prices in stratum ‘Helsinki, subarea 4, one-room’. Basic index Numbers from 2015.0 to 2020.4 (P(2015)=1). Block of flats.

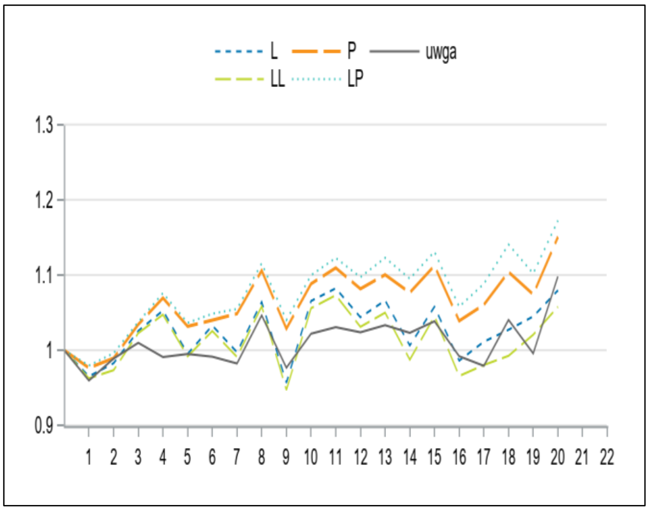
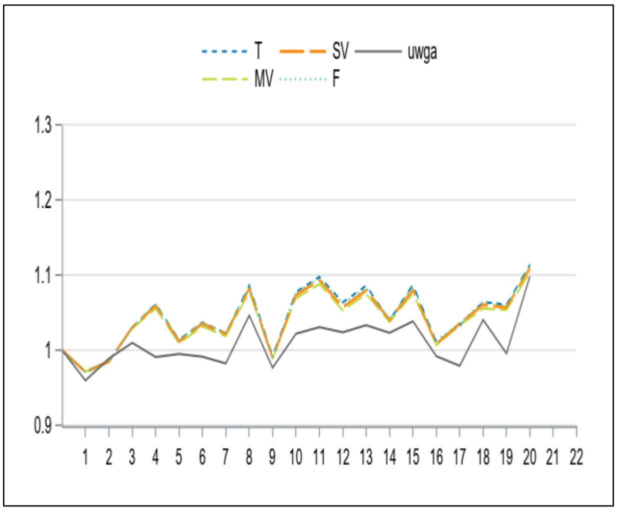


Figure 4: Hedonic index series for actual average prices in stratum ‘Helsinki, subarea 4, one-room’. Excellent index numbers from 2015.0 to 2020.4 (P(2015)=1). Block of flats.



Figures 1 to 4 present strategies for 1 and 3 in Table 4.3. In strategy 1 we aggregate observations by equal weights directly to conditional averages (i.e., unweighted geometric averages) for partition 1 and in strategy 3 by equal weights into partition 2 and then into partition 1 using index number formulae. We ask first whether weighting matters and second whether the index number formula matters. In the figures on the left we see that *basic index number formulae* with asymmetric weights are *contingently biased* (Vartia & Suoperä, 2017, 2018). Applying time antithesis for Laspeyres and rectification of Laspeyres by it we get Fishers ideal formula (*F*). Fisher is *excellent formula* (symmetric weights) and other three formulas (*T*, *MV* and *SV*) are quadratic approximation of Fisher – contradictory to basic formulas, excellent are very closely related. We conclude thus: Weighting matters and choice of formula matters – we prefer partition 2 and excellent index number formulas.

5.1.2 Synthesis of Weighted Arithmetic Average, WAA

Figure 5: Hedonic index series for actual average prices in stratum 'Espoo, subarea 1, three-rooms'. Basic index Numbers from 2015.0 to 2020.4 (P(2015)=1). Terraced houses.

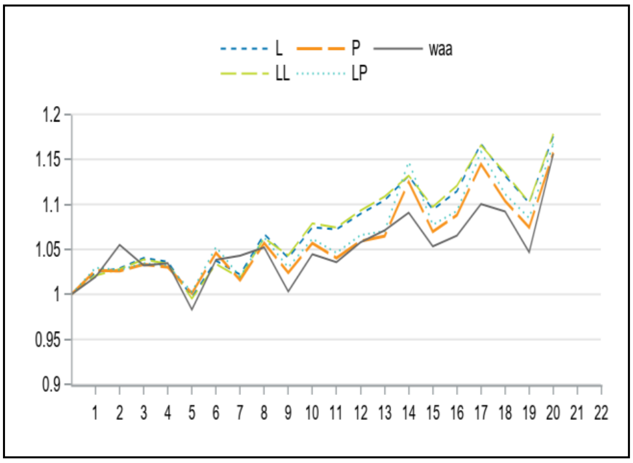


Figure 6: Hedonic index series for actual average prices in stratum 'Espoo, subarea 1, three-rooms'. Excellent index numbers from 2015.0 to 2020.4 (P(2015)=1). Terraced houses.

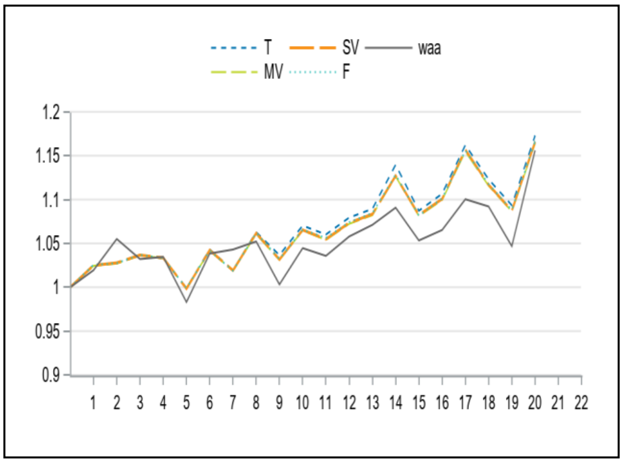


Figure 7: Hedonic index series for actual average prices in stratum 'Helsinki, subarea 4, one-room'. Basic index Numbers from 2015.0 to 2020.4 (P(2015)=1). Block of flats.

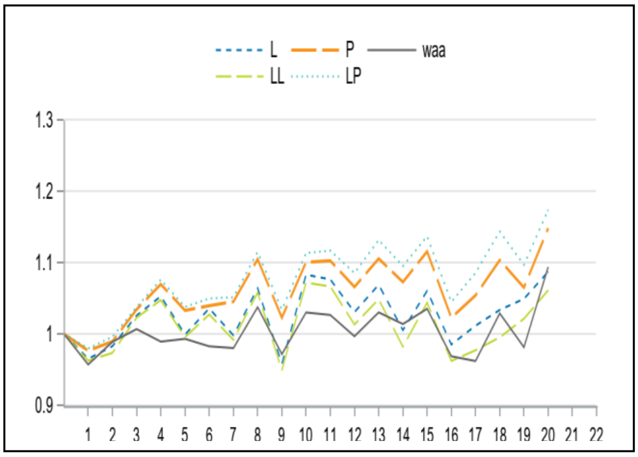
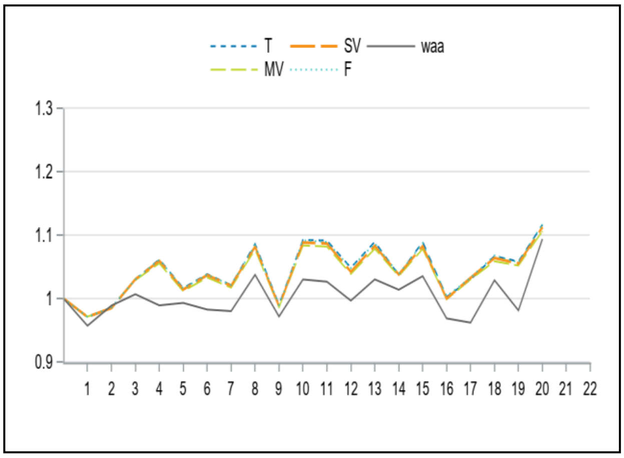


Figure 8: Hedonic index series for actual average prices in stratum 'Helsinki, subarea 4, one-room'. Excellent index numbers from 2015.0 to 2020.4 (P(2015)=1). Block of flats.



Figures 5 to 8 present strategies for 2 and 4 in Table 4.3. In strategy 2 we aggregate observations by weights in equation (4) directly to conditional weighted arithmetic averages for partition one and in strategy 3 by weights in equation (4) first into partition two and then into partition one using index number formulas. In left side figures we see that *basic index number formulas* with asymmetric weights are also here *contingently biased* (Vartia & Suoperä, 2017, 2018). In right-hand figures we see that *excellent formulas* with symmetric weights go hand-in-hand also here. In Figure 6 and 8 index series for conditional unit values (i.e., weighted arithmetic averages) are presented by solid lines (strategy 2). They both deviate from excellent index numbers (strategy 4) being seriously biased. The strategy 2 differs for two reasons: because of weighting and because of the inadequate quality adjustment due to regional price differences within postal areas and municipalities. Also, here as conclusion: Weighting matters and formula matters.

5.1.3 Synthesis Between Unweighted Geometric and Weighted Arithmetic Average

Statistical inference suggests using partition 2 instead of 1. Same holds also for hedonic quality adjusting – strategies based on partition 1 are not ideal. So, comparing strategies 3 and 4 is left. This is done by studying relations between index series constructed by base strategy of unweighted geometric and weighted arithmetic averages in strategies 3 and 4. We simply regress index series based on weighted arithmetic averages, say y, on index series based on unweighted geometric averages, say x. The index series used in regression are constructed by Törnqvist formula.

Model: $$y = a + \rho \cdot x + \varepsilon,$$

where y is index series constructed using weighted arithmetic average and x corresponding unweighted geometric average. Parameter ρ measures in this case correlation between y and x and results are

Table 4.4: Estimation results for above model (se in parenthesis).

Terraced houses	
Direct conditional averages	$\hat{\rho}$ = 0.96425 (0.00170), R^2 = 0.9629
Quality adjusted averages	$\hat{\rho}$ = 0.97088 (0.00170), R^2 = 0.9665
Block of flats	
Direct conditional averages	$\hat{\rho}$ = 0.98601 (0.00126), R^2 = 0.9674
Quality adjusted averages	$\hat{\rho}$ = 0.99456 (0.00133), R^2 = 0.9643

Correlation between y and x is almost 1 for all estimations. The following Figures for index series constructed by Törnqvist formula tell the same story – first, actual average price changes and second, corresponding quality adjusted price changes are closely related for conditional arithmetic and geometric averages.

Figure 9: Hedonic index series for actual average prices in stratum ‘Espoo, subarea 1, three-rooms’. Arithmetic solid line and geometric dotted from 2015.0 to 2020.4 (P(2015)=1). Terraced houses.

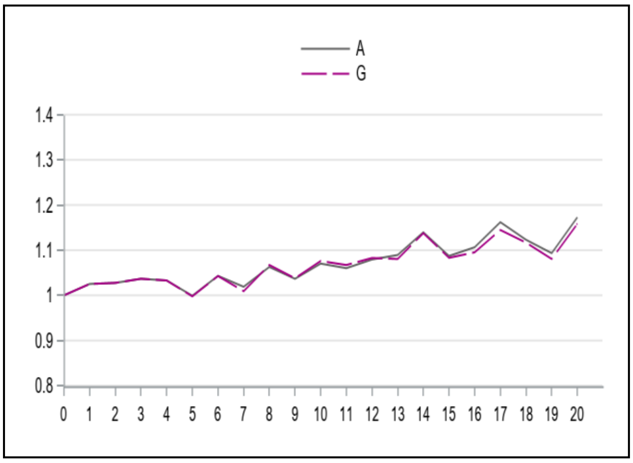


Figure 10: Hedonic index series for quality adjusted prices in stratum ‘Espoo, subarea 1, three-rooms’. Arithmetic solid line and geometric dotted from 2015.0 to 2020.4 (P(2015)=1). Terraced houses.

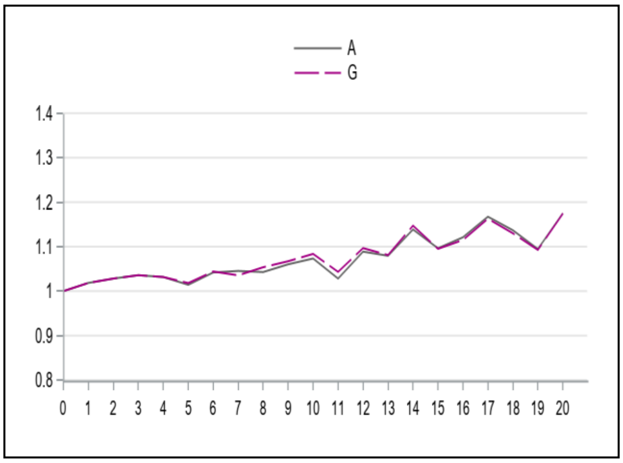


Figure 11: Hedonic index series for actual average prices in stratum ‘Helsinki, subarea 1, three-rooms’. Arithmetic solid line and geometric dotted from 2015.0 to 2020.4 (P(2015)=1). Block of flats.

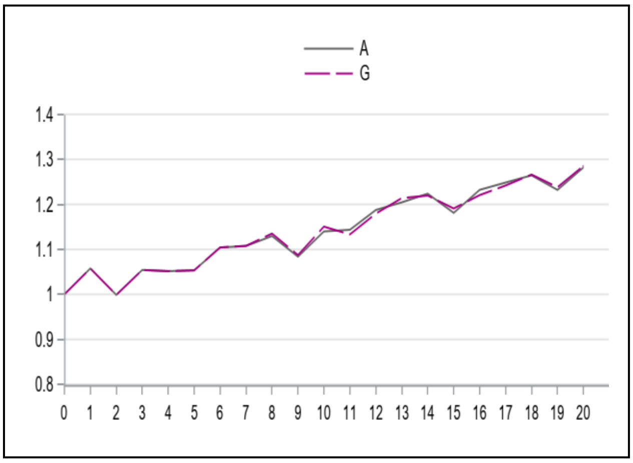


Figure 12: Hedonic index series for quality adjusted prices in stratum ‘Helsinki, subarea 1, three-rooms’. Arithmetic solid line and geometric dotted from 2015.0 to 2020.4 (P(2015)=1). Block of flats.

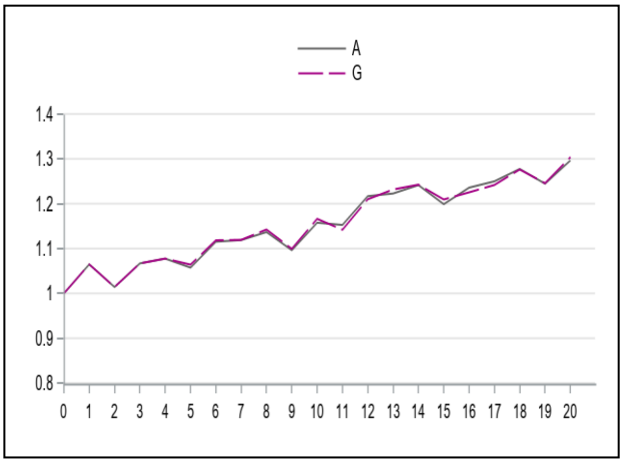


Table 4.4 tells that index series based on conditional averages (arithmetic and geometric) and their corresponding quality adjusted index series are almost identical for partition 2. Figures 9 to 12 complete that story – differences of hedonic index numbers constructed for weighted arithmetic and unweighted geometric (conditional) averages and their corresponding decompositions are of minimal significance for the most strata and their unions – either arithmetic or geometric may be selected as official statistics. Because Statistics Finland publishes also average price statistics, we suggest weighted arithmetic average for official production of house prices.

6 Conclusions

Instead of hedonic time-dummy approach, this study provides an alternative hedonic solution for quality adjusting. We combine FE-regression models (see Hsiao, 1986 p.29-32), well-known Oaxaca decomposition (Oaxaca, 1973) and traditional index number calculations, using transparent mathematics similarly as in Koev (2003) and Suoperä (2004, 2006). We do that for two alternative partitions, two price concepts – for unweighted geometric and weighted arithmetic averages – and for several basic and excellent index number formulas. Methods are always based on unbiased estimates of prices, quality characteristics and the BLUE of betas (homoscedastic errors) with aggregation rules of averages that are consistent in aggregation.

First, we make statistical inference of price models using our two nested partitions and get following results: 1. The price models are estimated very efficiently for terraced houses and block of flats (for all 34 equations). 2. Test statistics suggest using detailed partition based on postal areas (for more than 1000 stratums). 3. The price models are based on heterogeneously behaving cross-sections (significant heterogeneity component of behaviors, $He(x\beta)$) having detailed stratification (significant heterogeneity component of detailed stratification, $He(\alpha)$).

Second, we have knowledge of our estimated price models for which we aggregate into detailed stratum-level and make Oaxaca decomposition for them following the idea of Koev (2003), but here not only for unweighted geometric averages but also for weighted arithmetic averages (Suoperä, 2004, 2006). Both analyses are based on analysis of unbiased estimates. Similarly as in in statistical inference of price models we face the question of aggregation: should we aggregate observation level price models directly into partition one or two (more than 100 strata or more than 1000 strata) for calculating price ratios and decompositions of them. Both strategies are applied for unweighted geometric and weighted arithmetic average prices. We get the following results: 1. Direct aggregation from observations into strata of partition 1 using aggregation rule of unweighted geometric average leads to bias of unweighting (see eq. (8)). 2. Direct aggregation from observations into stratums of partition 1 using aggregation rule of weighted arithmetic average leads to bias of weighting (see eq. (9)). 3. The price models should aggregate into strata of partition 2. 4. Aggregation of price decompositions from strata in partition 2 using basic index number formulas with asymmetric weights leads to contingently biased index numbers for both unweighted geometric and weighted arithmetic averages. 5. Use excellent index number formulas in aggregation of stratums price decompositions all the time.

We suggest the following: 1. Use partition 2 in estimation of price models (for terraced houses and block of flats 34 equations including more than 1000 stratums). 2. Form price decompositions for stratums and aggregate them into 'crude' levels using excellent formulas, say for example Törnqvist. 3. Use aggregation rule of weighted arithmetic average, because of standard practice of publishing average prices.

This study give lessons about how '*weighted-by-economic-importance'* should be done using transparent algebra of unbiased estimates. As a warning: Never use asymmetric weighting as *weighted-by-economic-importance*-variable without control of quantity (or values) – here this is done effectively using accommodation-types in stratification.

References

Bailey M. J., Muth, R. F. and Nourse, H. O. 'A Regression Model for Real Estate Price Index Construction'., JASA, vol. 58, 933-942, 1963.

Case, K. E. and Shiller, R. J. 'Efficiency of the Market for Single Family Homes', American Economic Review, vol. 79, 125-137, 1989.

Davidson & MacKinnon 'Estimation and Inference in Econometrics', New York, Oxford University Press, 1993.

Diewert E. and Fox K. 'Substitution Bias in Multilateral Methods for CPI Construction using Scanner Data' 2018

Greene, W. "Econometric Analysis", Prentice-Hall International, Inc. (third ed.), 1997.

Hsiao, C. 'Analysis of Panel Data'., Cambridge University Press, 1986.

Koev, E. 'Combining Classification and Hedonic Quality Adjustment in Constructing a House Price Index', Licentiate thesis, Helsinki, 2003.

Koev, E. & Suoperä A. 'Pientalokiinteistöjen (omakotitalojen ja rakentamattomien pientalotonttien) hintaindeksit 1985=100', Helsinki, 2002. (in Finnish, Statistics Finland).

Oaxaca, R. 'Male-Female Wage Differentials in Urban Labour Markets', International Economic Review, 14, pp. 693-709, 1973.

Practical Guide on Multilateral Methods in the HICP (2020, WTPD-model), EuroStat.

Quigley, R. 'A Simply Hybrid Model for Estimating Real Estate Price Indexes', Journal of Housing Economics vol. 4, p. 1-12, 1995.

Rao, D.S. P. 'On the Equivalence of the Weighted Country Product Dummy (CPD) Method and the Rao System for Multilateral Price Comparisons', Review of Income and Wealth 51:4, 2005, 571-580.

Summers R. ‘International Comparisons with Incomplete Data", Review of Income and Wealth 29:1, 1973, pp. 1-16.

Suoperä, A. **’**Some new perspectives on price aggregation and hedonic index methods: Empirical application to rents of office and shop premises’, 2004, 2006 (unpublished, Statistics Finland).

Suoperä A. & Auno V. ‘Hedonic Index Numbers for Rents of Office and Shop Premises in Finland’, 2021, https://www.researchgate.net/publication/350460207_Hedonic_Index_Numbers_for_Rents_of_Office_and_Shop_Premises_in_Finland

Suoperä, A., Nieminen, K., Montonen, S. and Markkanen H. “Comparing Basic Averages, Index Numbers and Hedonic Methods as Price Change Statistic”, 2021, http://www.stat.fi/meta/menetelmakehitystyo/index_en.html)

Suoperä, A. & Vartia, Y. ‘Analysis and Synthesis of Wage Determination in Heterogeneous Cross-sections’, Discussion Paper No. 331, 2011.

Vartia, Y. & Suoperä, A. “Contingently biased, permanently biased and excellent index numbers for complete micro data”, 2018. (http://www.stat.fi/static/media/uploads/meta_en/menetelmakehitystyo/contingently_biased_vartia_suopera_updated.pdf)

Vartia, Y., Suoperä, A. and Vuorio, J. ’Hedonic Price Index Number for New Blocks of Flats and Terraced Houses in Finland’, 2021 (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html).

Vartia, Y. ‘Relative Changes and Index Numbers’, Ser. A4, Helsinki, Research Institute of Finnish Economy, 1976.

Vartia, Y. ‘Ideal Log-Change Index Numbers’, Scandinavian Journal of Statistics., 3, pp. 121-126,1976.

Vartia, Y. **’**Kvadraattisten mikroyhtälöiden aggregoinnista’, ETLA, Discussion Papers no. 25,1979.

Vartia, Y. & Suoperä, A. “Index number theory and construction of CPI for complete micro data”, 2017. (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html).

Vartia, Y. & Suoperä, A. “Contingently biased, permanently biased and excellent index numbers for complete micro data”, 2018. (http://www.stat.fi/static/media/uploads/meta_en/menetelmakehitystyo/contingently_biased_vartia_suopera_updated.pdf)

Vartia, Y. and Vartia, P. ‘Descriptive Index Number Theory and the Bank of Finland Currency Index’, Scandinavian Journal of Economics, vol. 3, pp. 352 . 364, 1985.

Törnqvist, L. ‘A Memorandum Concerning the Calculation of Bank of Finland Consumption Price Index’, unpublished memo, Bank of Finland, 1935.

Törnqvist, L. ’Levnadskostnadsindexerna i Finland och Sverige, Deras Tillförlitlighet och Jämförbarhet’, Ekonomiska Samfundets Tidskrift, vol. 37, 1-35, 1936.

Törnqvist, L. & Vartia, P. & Vartia, Y. ‘How Should Relative Changes be Measured’? The American Statistician, Vol. 39, No. 1. pp. 43 - 46, 1985.