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

Vartia, Yrjö, University of Helsinki
Suoperä, Antti, Statistics Finland
Vuorio, Johanna, Statistics Finland
2019

Abstract

Statistics Finland has relatively long experience in constructing indices of prices of old flats using the hedonic approach where regression analysis and classification of flats (i.e. stratification) are combined. Practically this means: First the price model is based on the fixed effect model applied for separate regions and second the quality adjustment is performed within each regional cell in a classification (i.e. inside stratums) using so called Oaxaca decomposition. The hedonic price index is computed by aggregating regional cell level (i.e. stratum level) quality adjusted prices using the logarithmic Laspeyres formula.

In this study we develop hedonic price index for new flats. In outline, we use similar methods as in case of old flats, but we develop Oaxaca decomposition in the cell or stratum level for unweighted and weighted arithmetic and geometric means. For that we develop new theorem for price aggregation from observation level to stratums by logarithmic mean. The new theorem of price aggregation is performed to semilogarithmic price models. The stratum level Oaxaca decomposition, that is, quality adjusted price changes and quality corrections, are aggregated using several basic (i.e. Laspeyres, Paasche, log-Laspeyres, log-Paasche,…) and excellent index number formulas (i.e. Törnqvist, Fisher, Mongomery-Vartia,…). The construction strategy of index series in this study is based on the base strategy. By this strategy index series are free of chain error (or drift).

Our test data is a high-quality register data on all free-market transactions of dwellings in new blocks of flats and terraced houses. Data includes statistical unit specific information of unit prices, quantities, values and of some unit specific quality characteristics form 2010/I to 2018/IV being quarter data. The quarter data includes about 2000 – 3000 observation per a quarter.

1 Introduction

Price index number using matched pairs method (see, Bailey, Muth and Nourse, 1963; Case and Shiller, 1989; Quigley, 1995) is not available for New Blocks of Flats and Terraced Houses, because each transacted flat appears only once in data. In this study, we use a method that calculates price changes for stratified cross-sectional samples of the data. The method combines relevant stratification, i.e. classification, of the studied topic on the one hand, and regression analysis of heterogeneous cross-sectional data, on the other. The index application of the method is based on the Oaxaca decomposition (Oaxaca, 1973), which breaks change in average prices (i.e. arithmetic and geometric) down into quality adjustment factors and price change standardized for quality.

Koev (2003) shows excellent analysis of hedonic price index number using Oaxaca decomposition of relative change for unweighted geometric average prices for semi-logarithmic price models. In this study, we develop three similar new theorems of price aggregation in which logarithmic prices at the observation level are aggregated to stratum level, so that we get three Oaxaca decomposition based on weighted or unweighted arithmetic or weighted geometric averages. The analysis of both arithmetic averages is based on the properties of logarithmic mean (see Törnqvist, 1935, p. 35; Y. Vartia, 1976, p.25; Törnqvist, P. Vartia and Y. Vartia, 1985, p. 44). 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 two we present notations. In chapter three we present the analysis of heterogeneous cross-sectional data, its stratification and estimation of unknown parameters of price models. Chapter four presents the price aggregation from observation level into strata. In chapter five we derive Oaxaca decomposition for different stratum level aggregates (i.e. unweighted and weighted arithmetic and geometric means). In chapter five we show how different index number formulas may be applied consistently for different stratum aggregates. Chapter six presents the results of the study and Chapter seven concludes.

The structure of study is based on three issues: 1. statistical inference of price models, 2. theory of price aggregation and 3. theory of hedonic price index numbers. In every part we use semi-logarithmic price model.

2 Basic Concepts and Notation

Our notation for the hedonic regression and for index number calculations is the following:

Commodities: $a_1, a_2, \ldots, a_{n_t}$ are transactions of dwellings in new blocks of flats and terraced houses in period *t*. The number n_t is about 2500.
Time periods: $t = 0, 1, 2, \ldots$ are the compared situations and are quarters.
Prices: p_{it} is the unit price of a_i in period *t.*
Quantities: q_{it} is the quantity of a_i in period *t.*
Explanatory variables in regressions: $\boldsymbol{x}_{it} = (x_{it1} \ldots x_{itk})'$ is a *k*-vector of observed characteristics in period *t.*
Values:$v_{it} = q_{it}p_{it}$ is the value of a_i in period *t.*
Total value: $V_t = \sum_i v_{it}$ is the total value of all the commodities in period *t*.
Total value ratio: $V^{t/0} = V^t/V^0$ is the total value ratio from period 0 to *t*.
Price relatives: $p_i^{t/0} = p_{it}/p_{i0}$ is the price relative of a_i from period 0 to *t*.
Quantity relatives: $q_i^{t/0} = q_{it}/q_{i0}$is the quantity relative of a_i from period 0 to *t*.
Value relatives: $v_i^{t/0} = v_{it}/v_{i0}$ is the value relative of a_i from period 0 to *t*.
Value shares: $w_{it} = v_{it}/\sum_i v_{it}$ is the value share of a_i in period *t.*

Corresponding n_t-vectors are denoted by the same symbol without sub-index of dwellings:

(1) $$\left(\boldsymbol{p}_t,\ \boldsymbol{q}_{t\,,}\boldsymbol{v}_t,\ \boldsymbol{w}_t,\ \boldsymbol{p}^{t/0},\ \boldsymbol{q}^{t/0},\ \boldsymbol{v}^{t/0}\right)$$

We assume that all prices and quantities are strictly positive (contain no zeros). This implies that all values, price, quantity and value relatives and value shares are also well-defined and strictly positive.

3 Analysis of Heterogeneous Cross-sectional Data

The analysis combines classification and typical regression analysis. In statistical terms, the method combines analysis of variance and typical regression analysis and is called as analysis of covariance model (se FE-model Hsiao, 1986 p.29-32).

3.1 Classification of Observations

We examine time periods $t = 0,1,\ldots,T$ and the finite set $A = \{a_1, a_2, \ldots, a_{n_t}\}$ of dwellings from new blocks of flats and terraced houses. *Each observation appears only once in the data and bilateral price links for homogeneous dwellings (matched pairs) cannot be applied.* Solution for this kind of cases is to form classification of observations into micro classes or stratum A_k, $k = 1,\ldots,K$, so that $A_k \cap A_r = \emptyset,\ \forall\ k \neq r$ and $A = \bigcup_{k=1}^{K} A_k$. In this study the micro partition is based on twelve separate regions which are divided into four different types of flats (one-room, two-rooms, three-rooms or bigger in block of flats and terraced houses). So, the partition includes 48 separate micro class at all time period *t.* The idea is to aggregate observations into strata level and then calculate some appropriate indicator for price change for each stratum.

As an example, we take one stratum A_k 'Two-rooms in Helsinki' and show basic problems of the study. Price determination of dwellings depends on at least location area (big towns like Helsinki, Tampere and Espoo, and

small municipals in NUTS2 regions), number of square meters of a dwelling, flat-type, distance from the center of municipal services and the owner of the building lot.

Figure 1: Price changes of unweighted and weighted geometric and arithmetic average prices in stratum 'Helsinki two-rooms' from 2010/1 – 2018/4[1].

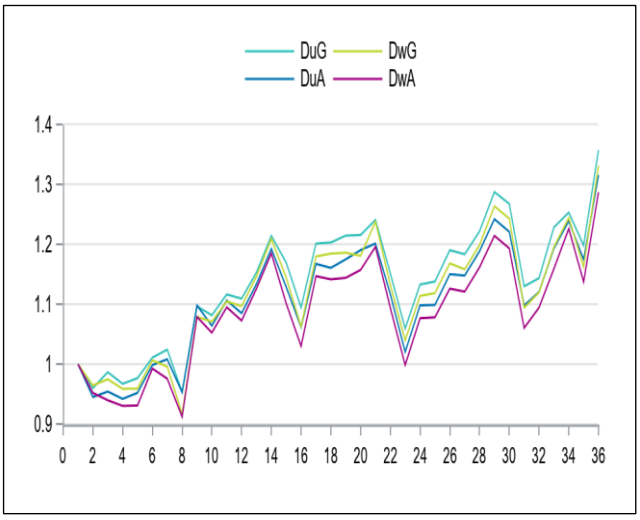
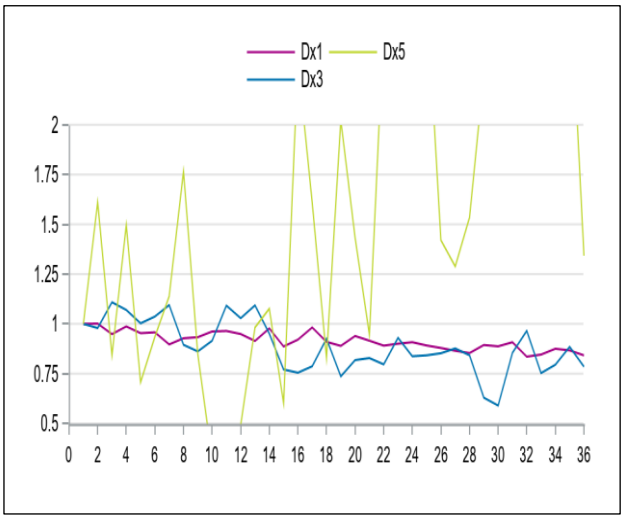


Figure 2: Changes of unweighted arithmetic average of quality characteristics in stratum 'Helsinki two-rooms' from 2010/1 – 2018/4[2].



All differences are calculated with respect to 2010/1 and quality characteristics are: x1, number of square meters of a dwelling, x3, distance from the center of municipal services and x5, owner of the building lot.

Figure one show actual or true price changes for unweighted and weighted arithmetic and geometric averages for stratum 'Two-rooms in Helsinki'. Figure two show how unweighted arithmetic averages for quality characteristics changes at the same time. Figures tell us that price changes for different aggregates differ a little the weighted arithmetic average being the lowest. At the same time the quality characteristics vary much stronger, especially "the owner of a building lot".

It seems that the number of square meters and distance declines from 2010 up to 2018 and a building lot is increasingly rented. These two figures tell that each aggregate is not comparable in quality. Especially when these quality characteristics have significant effects on prices, we need to remove the quality differences from price index numbers based on unweighted or weighted arithmetic and geometric averages. For controlling the quality change of characteristics, we need regression analysis.

3.2 Specification of the Price Model

Let's examine the data generating process of flat prices for a given region $r = 1, 2, \ldots 12$. Each region is stratified into four types of flats. The price equation for any region is specified as semi-logarithmic regression model, which is generally called to fixed-effects dummy-variable approach (Hsiao, 1986, s.29-32). We specify the model as linear in respect to parameters

$$\log(p_{irt}) = \alpha_{r1t} + \cdots + \alpha_{r4t} + \boldsymbol{x}'_{irt}\boldsymbol{\beta}_{rt} + \varepsilon_{irt} \quad (2)$$

[1] DuG = price change of unweighted geometric mean, DwG = price change of weighted geometric mean, DuA = price change of unweighted arithmetic mean, DwA = price change of weighted arithmetic mean,

[2] quality characteristics: x1 = number of square meters of a dwelling, x3= distance from the center of municipal services and x5= owner of the building lot.

where $\log(p_{irt})$ represents flat *i* specific logarithmic unit price per square meter in region *r* in period *t*. *k* vector of parameters $\boldsymbol{\beta}_{rt}$ in the regression model are allowed to vary according to regional grouping and time. Parameters $\alpha_{r1t}, \ldots \alpha_{r4t}$ represent flat type effects in the region *r* in period *t*. The *k* vector $\boldsymbol{x}'_{irt}$consists of exogenous independent variables (i.e. quality characteristics) typically used in empirical analysis of housing prices (number of square meters and its square root, distance by driving time to the gravitation point of different services like shopping center, municipal services and their square roots, owner of a building lot; own or rented). The equation (2) has non-linear square root terms in number of square meters and distance in driving times. Term ε_{irt} is random error term, which does not contain systematic information about the data generating process of flat prices. It is assumed, that $E\left(\varepsilon_{irt}\middle|\boldsymbol{x}'_{irt}\right) = 0$ and $Var\left(\varepsilon_{irt}\middle|\boldsymbol{x}'_{irt}\right) = \sigma_{rt}^2 < \infty$. In our model specification, the error covariance matrix is diagonal – a most natural situation for cross-sectional data.

3.3 Estimation of the Price Model

Estimation of unknown parameters follows the ordinary-least-squares (OLS) method, where - under the properties ε_{irt} – the OLS estimators are the best linear unbiased estimators (BLUE). The OLS estimators are obtained by minimizing the residual sum of squares. For apprehension of our study we use two step OLS method (see Davidson & MacKinnon, 1993, p. 19-25), where we transform observations as deviation of means with respect to our partition. Practically this means that each region (twelve separate region) are partitioned into four types of flats. So, in our case, the OLS estimator for $\boldsymbol{\beta}_{rt}$ is

(3) $$\widehat{\boldsymbol{\beta}}_{rt} = \left[\sum_{i=1}^{n_r}\sum_{k=1}^{4}(\boldsymbol{x}_{irkt} - \overline{\boldsymbol{x}}_{rkt})\,(\boldsymbol{x}_{irkt} - \overline{\boldsymbol{x}}_{rkt})'\right]^{-1}\sum_{i=1}^{n_r}\sum_{k=1}^{4}(\boldsymbol{x}_{irkt} - \overline{\boldsymbol{x}}_{rkt})\left(log(p_{irkt}) - log(\bar{p}_{krt}^{uG})\right)$$

where $log(p_{irkt})$ is observed logarithmic price per square meter for *i* belonging to strata *k* in region *r* in period *t* and the elements of vector $\boldsymbol{x}_{irkt}$ are corresponding exogenous explanatory variables (i.e. so-called quality characteristics). The term $log(\bar{p}_{krt}^{uG})$ is the arithmetic average of logarithmic flat prices (i.e. unweighted geometric average = *uG*) for strata *k* in region *r* in period *t* and elements of vector $\overline{\boldsymbol{x}}_{rkt}$ are arithmetic averages of explanatory variables in the same classes. The estimator is called also as the covariance estimator. In the second step partition-specific flat effects are estimated as

(4) $$\hat{\alpha}_{rkt} = log(\bar{p}_{krt}^{uG}) - \overline{\boldsymbol{x}}'_{rkt}\widehat{\boldsymbol{\beta}}_{rt}$$

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 is estimated by forcing the regression plane through the point of averages. This method is computationally extremely effective especially when partition includes thousands of strata (see Suoperä & Vartia, 2011).

The form of expression of the OLS estimators - simply log-prices, quality characteristics and their arithmetic and geometric averages - is useful for our new theorems of price aggregation. The basic axioms of the OLS method are well-known: 1) the OLS residuals sum to zero, 2) the regression hyperplane passes through the means of the data and 3) the average of fitted values from regression equals the average of the actual values of prices. These axioms will be satisfied for all strata.

We make distinction between the OLS estimation method and price aggregation – they may carry out separately, but such that for any price aggregates (i.e. unweighted or weighted arithmetic and geometric averages) these three axioms should be satisfied. The above unweighted geometric average is the simplest of them. Other three averages are unweighted arithmetic average and weighted arithmetic and geometric averages, which we derive in Chapter four.

3.4 Estimation Results of the Price Model

Our regression analysis for heterogeneously behaving cross-sections is standard statistical inference familiar to most statisticians, but how to present the multiplicity of the estimation results? We estimate twelve regional equations each having five explanatory variables and four strata according the flat type (block of flats: one-room, two-room, three-room or more and terraced houses). The number of estimates and their standard errors is 216, (that is 2*108) for each period. We have quarter data form nine years (2010-2018) meaning that just the number of estimates and their standard error is 1944. First time Suoperä (2004a, b, 2009a, b) and more elegantly Suoperä & Vartia (2011) show how the multiplicity of the estimation results for heterogeneously behaving cross-section may be presented. This approach is used again in this study.

We use equation (4) presented in Appendix 1 to show how effectively the heterogeneously behaving cross-section are estimated. First, we present the exogenous independent variables used in regressions and then the estimation results from eq. (4) in Appendix 1.

Table 1: The exogenous independent variables used in the regional price models

Independent variables	**Description of variable**
Flat type dummies	Classify observations into four flat types: one-room, two-room, three-room or more and terraced houses
x_1	Number of square meters of transacted flat
$x_2 = sqrt(x_1)$	Square root of the number of square meters of transacted flat
x_3	Distance from the center of municipal services
$x_4 = sqrt(x_3)$	Square root of the x_3 , distance from the center of municipal services
x_5	Owner of the building lot: Dummy variable that gets value 1, when the building lot is rented and otherwise 0.

Table 2: The estimation results for the price models in years 2017 and 2018 (t = 0 refers the whole year and other t = 1, 2, 3, 4 quarters of corresponding year).

Year	2017	2017	2017	2017	2017	2018	2018	2018	2018	2018
t	0	1	2	3	4	0	1	2	3	4
n_t	10759	2598	2762	2508	2891	9674	2418	2403	2395	2458
Strata	48	48	48	48	48	48	48	48	48	48
Adjr2	0.8112	0.8519	0.8381	0.8184	0.8177	0.8332	0.8500	0.8594	0.8231	0.8504
RMSE	0.1233	0.1168	0.1180	0.1124	0.1165	0.1204	0.1141	0.1123	0.1165	0.1174
$\hat{\alpha}_t$	9.9870	9.8321	9.9680	10.0868	10.05556	9.9357	10.0930	9.9067	9.8981	10.0690
se($\hat{\alpha}_t$)	(0.0247)	(0.0501)	(0.0455)	(0.0477)	(0.0460)	(0.0262)	(0.0529)	(0.0502)	(0.0538)	(0.0477)
$\hat{\beta}_{1t}$	0.01758	0.01581	0.01662	0.01873	0.01815	0.01731	0.01870	0.01656	0.01689	0.02055
se($\hat{\beta}_{1t}$)	(0.00044)	(0.00087)	(0.0008)	(0.00088)	(0.00080)	(0.00047)	(0.00094)	(0.00087)	(0.00095)	(0.00083)
$\hat{\beta}_{2t}$	-0.31838	-0.29166	-0.30485	-0.33571	-0.33186	-0.30682	-0.33913	-0.29004	-0.2978	-0.35709
se($\hat{\beta}_{2t}$)	(0.0067)	(0.0132)	(0.0123)	(0.0132)	(0.0120)	(0.0070)	(0.0142)	(0.0131)	(0.01424)	(0.01255)
$\hat{\beta}_{3t}$	0.013531	-0.0272	0.023516	0.023521	0.00595	0.015629	0.012956	0.026981	0.0173	-0.00983
se($\hat{\beta}_{3t}$)	(0.00059)	(0.00101)	(0.00093)	(0.00116)	(0.00311)	(0.00127)	(0.00153)	(0.00340)	(0.003838)	(0.00329)
$\hat{\beta}_{4t}$	-0.19646	-0.08807	-0.23375	-0.24617	-0.17212	-0.18717	-0.19314	-0.24502	-0.19053	-0.09306
se($\hat{\beta}_{4t}$)	(0.00321)	(0.0057)	(0.00566)	(0.00626)	(0.01256)	(0.00536)	(0.00739)	(0.01375)	(0.015477)	(0.01315)
$\hat{\beta}_{5t}$	-0.11114	-0.11919	-0.12025	-0.09502	-0.11853	-0.09873	-0.08909	-0.09675	-0.09184	-0.11705
se($\hat{\beta}_{5t}$)	(0.00278)	(0.00509)	(0.00500)	(0.00504)	(0.00492)	(0.00266)	(0.00496)	(0.00506)	(0.005252)	(0.00519)
γ_t	1	1	1	1	1	1	1	1	1	1
se(γ_t)	(0.00502)	(0.00861)	(0.00927)	(0.01066)	(0.01018)	(0.00510)	(0.00992)	(0.00929)	(0.010395)	(0.00989)
$\boldsymbol{\lambda}_t$	1	1	1	1	1	1	1	1	1	1
se($\boldsymbol{\lambda}_t$)	(0.0060)	(0.00888)	(0.01156)	(0.01194)	(0.01007)	(0.00572)	(0.01153)	(0.01020)	(0.010981)	(0.00975)

The estimation results are summarized as:

1. All parameters for representative behavior are statistically significant.
2. 'Flat-type' indicators (i.e. γ_t) for different regions are strongly significant and must be included into the price models
3. The data should be analyzed by heterogeneously behaving cross-section (i.e. time and regionally varying betas; in Table parameters $\boldsymbol{\lambda}_t$).
4. All quality characteristics have a negative effect on prices.

To interpret the estimation results of eq. (1) in Appendix 1 we take partial derivatives with respect to number of square meters of transacted flat (x_1) and driving time (x_3), that is

$$\frac{\partial \log(p_{irt})}{\partial x_{i1rt}} = \hat{\beta}_{1rt} + \hat{\beta}_{2rt}/sqrt(x_{i1rt}), \forall\, i \in A_r \text{ and}$$
$$\frac{\partial \log(p_{irt})}{\partial x_{i3rt}} = \hat{\beta}_{3rt} + \hat{\beta}_{4rt}/sqrt(x_{i3rt}), \forall\, i \in A_r.$$

When we calculate cumulative sums of partial derivates from ordered samples (i.e. x_{i1}, x_{i3} are ordered starting from smallest), we get Figures 3 and 4. In Figure 3 we see that square meter prices fall when number of squares increase. The results are approximately equal compared to Koev (2003, p. 34) – square meter price of a 60 m^2 flat is about 30 – 35 log-% lower compared to a 20 m^2 flat.

Figure 3: The effect of size on the square meter of apartment (periods, 2017.m, 2018.m, m = 0,1,2,3,4)

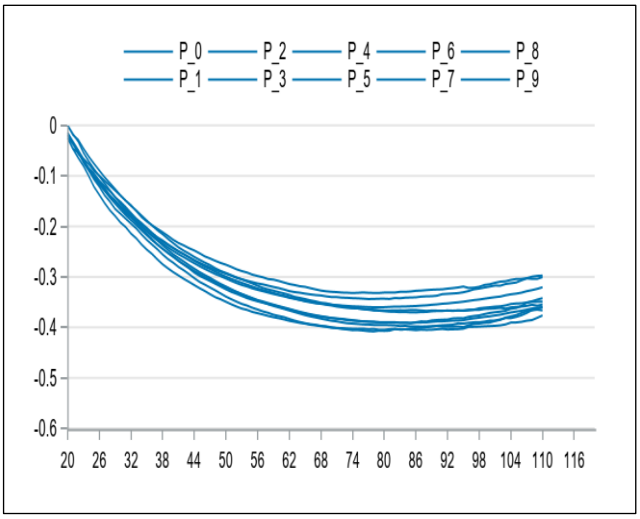
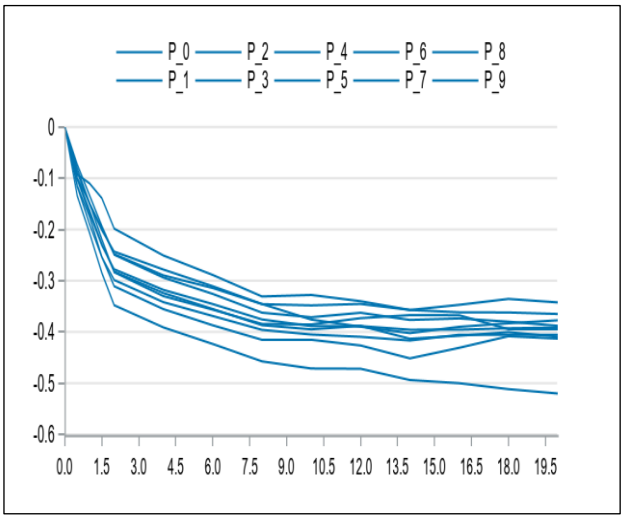


Figure 4: The effect of distance on prices of apartment (periods, 2017.m, 2018.m, m = 0,1,2,3,4)



The effect of 'distance' on prices is also significant. Prices decline approximately 40 log-% after 10-kilometer distance. The 'distance'-variable is very important explanatory variable in determination of flat prices in Finland.

Next, we examine the price aggregation of estimated price models (1) (see Appendix 1) from the observation into aggregate level (i.e. strata level).

4 Price Aggregation from Observations into Stratum Aggregates

The most common method for price aggregation from observation into strata level (i.e. elementary aggregate level) is to choose equal weights for all observations. In this case the semi-logarithmic specification of estimated model (2) leads to an unweighted geometric average price (*uG*) at left side of (2) and at the right side of (2) to unweighted arithmetic average (*uA*) of independent explanatory variables or quality characteristics. The method is based on typical aggregation of unit prices in which all observations have the same weight and therefore contribute equally to stratum averages irrespective of their real quantitative categories. This is also our first method of price aggregation – subsequent three (unweighted arithmetic, and weighted arithmetic and geometric means) are more complicated.

First, we define logarithmic mean for two positive numbers *x* and *y* as follows (L. Törnqvist, 1935, p. 35; Y. Vartia, 1976; L. Törnqvist, P. Vartia and Y. Vartia, 1985, p. 44)

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

The definition can also be expressed as $\log(y/x) = (y-x)/L(x,y)$, when it connotes that the log change is a relative change in respect to the logarithmic mean. This indicator of relative change is a ratio that is symmetrical, additive and independent of measurement unit.

Let us examine positive sets of numbers $\{x_1, x_2, \dots, x_n\}$ and $\{y_1, y_2, \dots, y_n\}$ and define their logarithmic mean

$$L(\sum_i^n y_i, \sum_i^n x_i) = \frac{\sum_i^n y_i - \sum_i^n x_i}{\log(\sum_i^n y_i / \sum_i^n x_i)}, \text{ or}$$

$$\log(\sum_i^n y_i / \sum_i^n x_i) = \sum_i^n \frac{y_i - x_i}{L(\sum_i y_i, \sum_i x_i)}$$

and by definition of logarithmic mean the above equation reduces to

(6) $$\log(\sum_i^n y_i / \sum_i^n x_i) = \sum_i \frac{L(y_i, x_i)}{L(\sum_i y_i, \sum_i x_i)} log(y_i / x_i)$$

This equation is just what we need for aggregation of observed prices into arithmetic averages for stratum aggregates (Suoperä, 2006, p.4). The operational characteristics of (6) are so far difficult to see. If we concretize the expression by choosing: $y_i = v_i = q_i p_i$ (i.e. values) and $x_i = q_i$ (i.e. quantities), we get

(7) $$\log(\sum_i^n v_i / \sum_i^n q_i) = log(\bar{p}^{wA}) = \sum_i \frac{L(v_i, q_i)}{L(\sum_i v_i, \sum_i q_i)} log(p_i)$$

This is our new theorem of price aggregation – we see, that the argument of log-function in left side of (7) is the weighted arithmetic average (*wA*) of observed prices. When we choose: $y_i = v_i = q_i p_i$ and $x_i = q_i = 1$, we get

(8) $$\log(\sum_i^n p_i / \sum_i^n q_i) = log(n \cdot \bar{p}^{uA} / n) = \log(\bar{p}^{uA}) = \sum_i \frac{L(p_i, 1)}{L(\sum_i p_i, n)} log(p_i)$$

Now the argument of log-function is the unweighted arithmetic average (*uA*). Both principles of price aggregation have used in official production of statistics for rents of office and shop premises since 2002 (Suoperä, 2002, 2006). In conclusion, we put together weights used in our price aggregation.

Table 3: Weights for different elementary aggregates in price aggregation for semi-logarithmic price models.

Strata aggregate	Mathematical formula for weights
Unweighted arithmetic average (*uA*)	$w_i^{uA} = \frac{L(p_i, 1)}{L(\sum_i p_i, n)}$
Weighted arithmetic average (*wA*)	$w_i^{wA} = \frac{L(v_i, q_i)}{L(\sum_i v_i, \sum_i q_i)}$
Unweighted geometric average (*uG*)	$w_i^{uG} = \frac{1}{n}, \forall i$
Weighted geometric average (*wG*)	$w_i^{wG} = \frac{q_i}{\sum_i q_i}$

4.1 Price Aggregation for the Semi-logarithmic Price Model

The aggregation of all variables of model (1) in Appendix 1 is guided by the three following mathematical characteristics: First, we demand that residuals for any single stratum sum up to zero. Second, the hypersurface of the regression model must always go through the averages of dependent and exogenous independent explanatory variables. Third, the mean of the fitted values, i.e. predictions of log-prices, must match precisely the arithmetic average of observed log-prices (i.e. OLS solution). *These three axioms reveal that the dependent variable is decomposed into two orthogonal components of which the first is expressed as a linear combination of exogenous variables and the second as an error term (residual), which is orthogonal with the exogenous variables of the model*. Practically this means that the OLS method satisfy all above properties first time at the strata level. We simplify our analysis to one stratum A_k, that is a subset of region *r* in time period *t*. Then we apply the above method explained in italics to aggregation of eq. (1) in Appendix 1. First, we define most simply weights, $w_{ikt} = 1/n_t$ (i.e. equal weights for all *i*), and aggregate estimated equation into stratum A_k, that is (note that in eq. (1), $\widehat{\boldsymbol{\beta}}_{kt} = \widehat{\boldsymbol{\beta}}_{rt}, \forall k \in r$)

$$\sum_i w_{ikt} \log(p_{ikt}) = \sum_i w_{ikt} \left(\hat{\alpha}_{kt} + \boldsymbol{x}'_{ikt}\widehat{\boldsymbol{\beta}}_{kt}\right) + \sum_i w_{ikt}\, \hat{\varepsilon}_{ikt} \tag{9}$$

For semi-logarithmic price model the estimate for 'flat-type' indicator in stratum $A_k$is $\hat{\alpha}_{kt} = log\left(\bar{p}_{kt}^{uG}\right) - \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{kt}$, where $\bar{p}_{kt}^{uG}$ is unweighted geometric average of prices (*uG*) and $\overline{\boldsymbol{x}}'_{kt}$ is arithmetic averages of explanatory variables (*uA*). As we see, the right side of equation is divided into two parts which in the case of the OLS method may expressed as

$$\sum_i w_{ikt} \left(\hat{\alpha}_{kt} + \boldsymbol{x}'_{ikt}\widehat{\boldsymbol{\beta}}_{kt}\right) = \sum_i w_{ikt} log\left(\bar{p}_{kt}^{uG}\right) - \sum_i w_{ikt}\, \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{kt} + \sum_i w_{ikt}\, \boldsymbol{x}'_{ikt}\widehat{\boldsymbol{\beta}}_{kt} \tag{10a}$$

$$\sum_i w_{ikt}\, \hat{\varepsilon}_{ikt} = \sum_i w_{ikt} \log(p_{ikt}) - \sum_i w_{ikt} log\left(\bar{p}_{kt}^{uG}\right) + \sum_i w_{ikt}\, \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{kt} - \sum_i w_{ikt}\, \boldsymbol{x}'_{ikt}\widehat{\boldsymbol{\beta}}_{ikt} \tag{10b}$$

Next, we study suitable options for weights w_{ikt} (see, Table 3). After simple algebra we get following stratum aggregates represented in Table 4.

Table 4: Price aggregation from observations into stratum level for a semi-logarithmic model estimated by the OLS.

Statistics	Mathematical formula for weights w_{ikt}, $\forall i \in A_k$	Stratum aggregate
Unweighted arithmetic average (*uA*)	$w_{ikt}^{uA} = \dfrac{L(p_{ikt}, 1)}{L(\sum_i p_{ikt}, n)}$	$\log\left(\bar{p}_{kt}^{uA}\right) = \hat{\alpha}_{kt}^{uA} + \overline{\boldsymbol{x}}'^{uA}_{kt}\widehat{\boldsymbol{\beta}}_{kt}$, where $\hat{\alpha}_{kt}^{uA} = log\left(\bar{p}_{kt}^{uA}\right) - \overline{\boldsymbol{x}}'^{uA}_{kt}\widehat{\boldsymbol{\beta}}_{kt}$
Weighted arithmetic average (*wA*)	$w_{ikt}^{wA} = \dfrac{L(v_{ikt}, q_{ikt})}{L(\sum_i v_{ikt}, \sum_i q_{ikt})}$	$\log\left(\bar{p}_{kt}^{wA}\right) = \hat{\alpha}_{kt}^{wA} + \overline{\boldsymbol{x}}'^{wA}_{kt}\widehat{\boldsymbol{\beta}}_{kt}$, where $\hat{\alpha}_{kt}^{wA} = log\left(\bar{p}_{kt}^{wA}\right) - \overline{\boldsymbol{x}}'^{wA}_{kt}\widehat{\boldsymbol{\beta}}_{kt}$
Unweighted geometric average (*uG*)	$w_{ikt}^{uG} = \dfrac{1}{n}$	$\log\left(\bar{p}_{kt}^{uG}\right) = \hat{\alpha}_{kt} + \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{kt}$, where $\hat{\alpha}_{kt} = log\left(\bar{p}_{kt}^{uG}\right) - \overline{\boldsymbol{x}}'_{kt}\widehat{\boldsymbol{\beta}}_{kt}$
Weighted geometric average (*wG*)	$w_{ikt}^{wG} = \dfrac{q_{ikt}}{\sum_i q_{ikt}}$	$\log\left(\bar{p}_{kt}^{wG}\right) = \hat{\alpha}_{kt}^{wG} + \overline{\boldsymbol{x}}'^{wG}_{kt}\widehat{\boldsymbol{\beta}}_{kt}$, where $\hat{\alpha}_{kt}^{wG} = log\left(\bar{p}_{kt}^{wG}\right) - \overline{\boldsymbol{x}}'^{wG}_{kt}\widehat{\boldsymbol{\beta}}_{kt}$

Table 4 defines our four stratum aggregates that are derived for the semi-logarithmic price models estimated by the OLS. As we see in Table 4, the standard textbook selection of weights gives standard OLS results at the stratum level (i.e. unweighted geometric average in above Table). All other three elementary aggregates are derived from the OLS solution. They have exactly similar form and are in fact reparametrized OLS solutions for

unweighted arithmetic, weighted arithmetic and weighted geometric averages. They satisfy three basic axioms of the OLS method similarly as standard textbook OLS method, that is: 1. residuals vanish, 2. the hypersurface of the regression model goes through averages of the variables and 3. averages of predicted log-prices of the model match precisely the corresponding averages calculated from actual log-prices.

Table 4 defines all that is necessary for construction of hedonic price index numbers for stratum $A_k, k = 1, \dots, K$. Next, we apply so called Oaxaca decomposition for different stratum aggregates to get strata level 'micro' indices for actual price ratios, quality changes and quality adjusted price ratios.

5 Hedonic Price Index Numbers and its Decomposition

In this study, a hedonic price index depends on 1. data in question (quality adjusting), 2. the strategy used (base, chain or rather a mixture of them), 3. index number formula and 4. aggregate used in calculation (i.e. $\log(\bar{p}_{kt}^{m}) = \hat{\alpha}_{kt}^{m} + \bar{\boldsymbol{x}}'^{m}_{kt}\hat{\boldsymbol{\beta}}_{kt}$, for all $m = uA, wA, uG, wG$, and $t = 0, 1, \dots$ see Table 4).

The key problem of the chain type indices (i.e. strategy) is the *chain error* (or chain drift) that tends to grow when chaining is applied frequently – typically on a monthly or quarterly basis. So, because chain error is *data contingent* and realized only for a chain-type strategy, we abandon the chain and favor the base strategy. In this study we use the base strategy, where the base period is defined to be previous year, which is normalized as average quarter. In fact, we use similar strategy as in Finnish CPI applied to a scanner type complete micro data (see, Vartia, Suoperä, Nieminen & Montonen, 2018a, 2018b; Vartia, Suoperä, Nieminen & Markkanen, 2019).

We analyze two sets of index number formulas. The first set is based on formulas using old or new 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 formula: Montgomery-Vartia (*MV*), Törnqvist (*T*), Fisher (*F*) and Sato-Vartia (*SV*). We call these index number formulas as *excellent formula*. For the fundamental analysis of these index number formula, see Vartia & Suoperä (2018).

Normally index number theory is not applied to hedonic methods. Koev & Suoperä (2002), Koev (2003) and Suoperä (2002, 2006, 2010a, 2010b) make an exception for that and include index number theory to hedonic method to get hedonic price index numbers. We continue in this study alike and use typical notation familiar to index number theory (Vartia & Suoperä, 2017, 2018).

5.1 Within Stratum Hedonic Quality Adjustment and its Decomposition

We have all information necessary to define so called Oaxaca decomposition for our four aggregates presented in Table 4. For simplicity, we have two time periods, the base ($t = 0$, previous year) and the observation quarter of current year ($t = 1$) and only one stratum A_k. We calculate difference between these two price models (0, *t*) separately for each stratum aggregate indexed by subindex $m = uA, wA, uG, wG$ (see Table 4), that is

$$\log(\bar{p}_{k1}^{m}/\bar{p}_{k0}^{m}) = \hat{\alpha}_{kt}^{m} + \bar{\boldsymbol{x}}'^{m}_{kt}\hat{\boldsymbol{\beta}}_{kt} - \hat{\alpha}_{k0}^{E} - \bar{\boldsymbol{x}}'^{m}_{k0}\hat{\boldsymbol{\beta}}_{k0}$$

Defining first Oaxaca decomposition (1973) and then exp-transformation, we get

$$\bar{p}_{k1}^{m}/\bar{p}_{k0}^{m} = exp\{(\bar{\boldsymbol{x}}'^{m}_{k1} - \bar{\boldsymbol{x}}'^{m}_{k0})\hat{\boldsymbol{\beta}}_{k1}\} \cdot exp\{\hat{\alpha}_{k1}^{m} - \hat{\alpha}_{k0}^{m} + \bar{\boldsymbol{x}}'^{m}_{k0}(\hat{\boldsymbol{\beta}}_{kt} - \hat{\boldsymbol{\beta}}_{k0})\} \quad (11)$$

The left side in (11) is simply the price ratio for a given stratum aggregate *m*. We call this price ratio as 'actual or true price change' (= *A*) for aggregate *m* (see Table 4). The first term on right side is a 'price change due to quality difference' (= *QC*) of the sample mix at current year observation period valuation of the characteristics (see Table 1). The second term on right is a 'quality adjusted price change' (= *QA*) evaluated at *standard point of quality*, that is $\bar{\boldsymbol{x}}'^{m}_{k0}$.

Index number formula is not needed for compilation of price change for stratum A_k– the direct price link from period 0 to period *t* is based purely on 'average statistics' of equation (11). The left side of (11) simply tells that average prices in base 0 and observation period *t* depends on corresponding averages of quality characteristics. When these vector of averages $(\bar{\boldsymbol{x}}^0, \bar{\boldsymbol{x}}^t)$ are not equal then the price ratio is not based on commodities that are comparable in quality. The price ratio of averages of 'actual or true prices', that is $\bar{p}_A^{t/0} = \bar{p}^t(\bar{\boldsymbol{x}}^t)/\bar{p}^0(\bar{\boldsymbol{x}}^0)$ is not satisfactory choice for official price change statistics – the quality differences should be removed from it. We have two main choice for the standard quality point of quality characteristics – either $\bar{\boldsymbol{x}}^0$ or $\bar{\boldsymbol{x}}^t$. The base period standard quality point $\bar{\boldsymbol{x}}^0$ is estimated from previous year and $\bar{\boldsymbol{x}}^t$ from quarter of current year. Previous year (i.e. period 0) includes about fourfold observations compared to observation quarter and so statistical properties of $\bar{\boldsymbol{x}}^0$ favor to select it as standard quality point. The following table collect together all information from eq. (11) in terms of prices.

Table 5: Components of (11) in price terms for any stratum aggregate *m* separately.

	Price ratio	Explicit formula
Actual or true price change, $p_A^{t/0}$	$\bar{p}^t(\bar{\boldsymbol{x}}^t)/\bar{p}^0(\bar{\boldsymbol{x}}^0)$	$exp\{(\hat{\alpha}_t + \bar{\boldsymbol{x}}_t'\hat{\boldsymbol{\beta}}_t) - (\hat{\alpha}_0 + \bar{\boldsymbol{x}}_0'\hat{\boldsymbol{\beta}}_0)\}$
Quality adjusted price change, $p_{QA}^{t/0}$	$\bar{p}^t(\bar{\boldsymbol{x}}^0)/\bar{p}^0(\bar{\boldsymbol{x}}^0)$	$exp\{(\hat{\alpha}_t + \bar{\boldsymbol{x}}_0'\hat{\boldsymbol{\beta}}_t) - (\hat{\alpha}_0 + \bar{\boldsymbol{x}}_0'\hat{\boldsymbol{\beta}}_0)\}$
Price change for quality correction of square meter, $p_{QC,x_1}^{t/0}$	$\bar{p}_{x_1}^t(\bar{x}_1^t, \bar{x}_2^t)/\bar{p}_{x_1}^0(\bar{x}_1^0, \bar{x}_2^0)$	$exp\{(\bar{x}_{1t}\hat{\beta}_{1t} + \bar{x}_{2t}\hat{\beta}_{2t}) - (\bar{x}_{10}\hat{\beta}_{1t} + \bar{x}_{20}\hat{\beta}_{2t})\}$
Price change for quality correction of distance, $p_{QC,x_3}^{t/0}$	$\bar{p}_{x_3}^t(\bar{x}_3^t, \bar{x}_4^t)/\bar{p}_{x_3}^0(\bar{x}_3^0, \bar{x}_4^0))$	$exp\{(\bar{x}_{3t}\hat{\beta}_{3t} + \bar{x}_{4t}\hat{\beta}_{4t}) - (\bar{x}_{30}\hat{\beta}_{3t} + \bar{x}_{40}\hat{\beta}_{4t})\}$
Price change for quality correction of owner of building lot, $p_{QC,x_5}^{t/0}$	$\bar{p}_{x_5}^t(\bar{x}_5^t)/\bar{p}_{x_5}^0(\bar{x}_5^0)$	$exp\{(\bar{x}_{5t}\hat{\beta}_{5t} - \bar{x}_{50}\hat{\beta}_{5t})\}$

Now the equation (11) may expressed by price ratios, that is

$$p_A^{t/0} \equiv p_{QC,x_1}^{t/0} \cdot p_{QC,x_3}^{t/0} \cdot p_{QC,x_5}^{t/0} \cdot p_{QA}^{t/0} \tag{12}$$

Each price ratio in (12) is estimated separately for a given price aggregate (i.e. $m = uA, wA, uG, wG$). Each x-variable - size of flat in square meters, distance and share of owner of a building lot - have negative effect on prices. When these x-variables are smaller than their standard quality point (i.e. $\bar{\boldsymbol{x}}^t - \bar{\boldsymbol{x}}^0 < \mathbf{0}$), all quality corrections $p_{QC,x_1}^{t/0}$, $p_{QC,x_3}^{t/0}$, $p_{QC,x_5}^{t/0} > 1$. This means that we need to adjust actual price change $p_A^{t/0}$ just amount of quality corrections $p_{QC,x}^{t/0} = p_{QC,x_1}^{t/0} \cdot p_{QC,x_3}^{t/0} \cdot p_{QC,x_5}^{t/0}$ downward to get quality adjusted price change $p_{QA}^{t/0}$ which is standardized of quality.

We have developed in addition to unweighted geometric average (that is a standard textbook method for semi-logarithmic heterogeneously behaving cross-sections) three other aggregates and their index solution in the stratum level. Next, we present the hedonic price index numbers by traditional notation of index number theory.

5.2 Hedonic Price Index Numbers

The equation (12) simply says that 'actual or true price change' (i.e. sub index *A*) of a given aggregate is divided in our case into four components – three of them are due to quality change (i.e. sub index *QC*) and last one is 'quality adjusted price change' (= *QA*). Equations (11) and (12) are identities for any aggregation level for given index number formula and aggregate. For simplicity we define our index number to price-link from period 0 to period 1. Replacing period 1 by *t* we get our base strategy familiar to our CPI analysis based on scanner-type complete micro data (see for example Vartia, Suoperä, Nieminen & Montonen, 2018a, 2018b). In this strategy the base period is previous year normalized as average quarter and observation period *t* is a quarter of current year. This is most natural choice, because it is free of chain error (or drift).

We analyze two sets of index number formulas. The first set is based on formulas using old (Laspeyres *L* and Log-Laspeyres *l*) or new weights (Log-Paasche *p* and Paasche *P*). These index number formulas are called as *basic formulas*. These formulas are data contingently biased and should be used. The second set of index numbers include four formulas: Montgomery-Vartia, *MV*, Törnqvist, *T*, Fisher, *F* and Sato-Vartia, *SV*). We call these index number formulas as *excellent formulas*. The fundamental analysis of these index number formulas, see Vartia & Suoperä (2018).

In Table 6 we collect together all information that is necessary for calculation of price indices. Table 6 shows that all index number formulas are presented in multiplicative form, including Laspeyres and Paasche, which are derived from its logarithmic representations (see Vartia, 1976, p.128). Practically this means, that aggregation of price changes in (12) is done always much simpler in logarithmic form (i.e. additive form) and then transformed back as indices. As we have stressed, all formulas are evaluated separately for each component of (12) and separately for aggregate in question.

Table 6: Necessary information for calculation of hedonic price indices for different formulas and aggregates (Vartia & Suoperä, 2017, 2018).

Basic formula: Contingently biased index numbers		
Symbol for the formula	$P^{1/0}$	Weights of the formula, w_i
L	$\prod\left(p_k^1/p_k^0\right)^{w_k^0}$	$w_k^0 = \frac{L(p_k^1 q_k^0, p_k^0 q_k^0)}{L(p^1 q^0, p^0 q^0)}$
l	$\prod\left(p_k^1/p_k^0\right)^{w_k^0}$	$w_k^0 = \frac{v_k^0}{V^0}$
p	$\prod\left(p_k^1/p_k^0\right)^{w_k^1}$	$w_k^1 = \frac{v_k^1}{V^1}$
P	$\prod\left(p_k^1/p_k^0\right)^{w_k^1}$	$w_k^1 = \frac{L(p_k^1 q_k^1, p_k^0 q_k^1)}{L(p^1 q^1, p^0 q^1)}$
Excellent formulas		
T	$\prod\left(p_k^1/p_k^0\right)^{\bar{w}_k}$	$\bar{w}_k = 0.5 * (w_k^0 + w_k^1)$
SV	$\prod\left(p_k^1/p_k^0\right)^{\bar{w}_k}$	$\bar{w}_k = \frac{L(w_k^1, w_k^0)}{\sum L(w_k^1, w_k^0)}$
MV	$\prod\left(p_k^1/p_k^0\right)^{\bar{w}_k}$	$\bar{w}_k = \frac{L(v_k^1, v_k^0)}{L(V^1, V^0)}$
F	$\left(L^{1/0} * P^{1/0}\right)^{1/2}$	

These index number formulas are used when strata decompositions are aggregated into crude aggregates like ‘flat-type’ -level in Finland etc. Aggregation of decomposition (12) are done in logarithmic form (i.e. in additive form) keeping stratum aggregate and index number formula fixed.

6 Empirical Results

6.1 Comparation of Stratum Statistics and Their Decomposition of ‘True or Actual Price Changes’

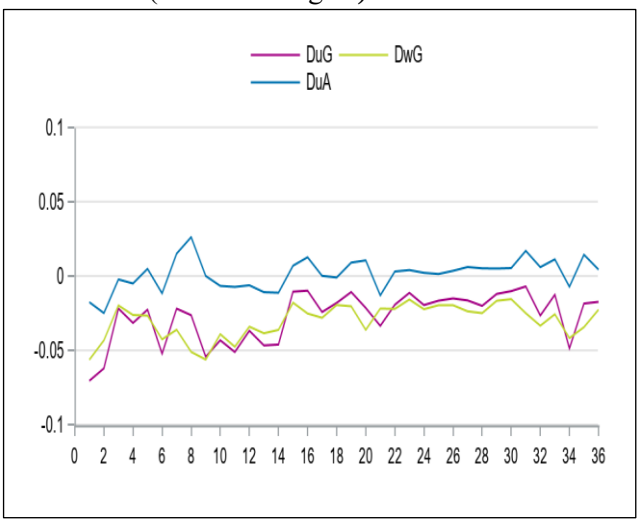
Our partition of transacted flats consists of 48 stratums (12 regions and every single one is divided into four flat-type). For each stratum we calculate four types of aggregates (see Table 4), for which holds

1. unweighted arithmetic average, $uA \geq uG$, unweighted geometric average
2. weighted arithmetic average, $wA \geq wG$, weighted geometric average

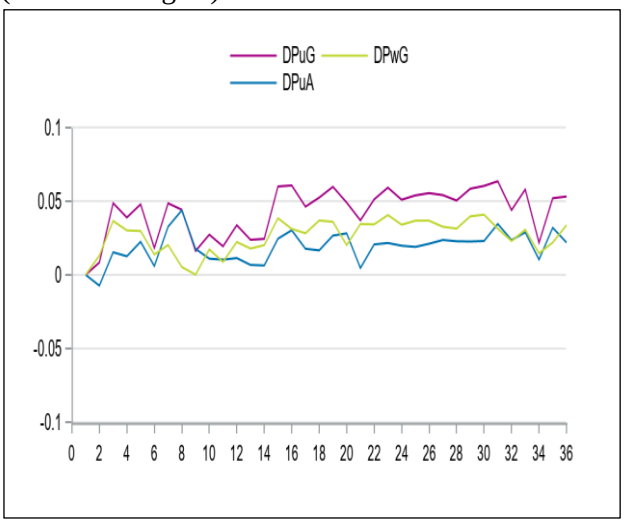
As an example, the following figures shows for stratum ‘Helsinki, two-rooms’ how significantly these aggregates deviate.

The left-side figure presents deviation between unweighted arithmetic (DuA=log(uA/wA)) and geometric ((DuG=log(uG/wA))) and weighted geometric averages ((DwG=log(wG/wA))) from weighted arithmetic averages (wA) in log-%. Here the largest deviation is about 7 log-%, but in some other strata it may be over 10 log-%. The right-side figure present how much index series of weighted arithmetic average deviates from other aggregates in log-scale. Here almost as a rule the index series of geometric averages exceeds both index series of arithmetic averages.

Figure 5: Ratio of averages with respect to weighted arithmetic average in ‘Helsinki, two-rooms’ (deviations log-%).



Fiqure 6: Ratio of index series of averages with respect to weighted arithmetic average in ‘Helsinki, two-rooms’ (deviations log-%).



For construction of the official average price statistics difference between these four aggregates are unexpected large especially for geometric averages (yellow and red lines). Quite unexpectedly the index series of geometric averages exceeds the weighted arithmetic average about 4 – 5 log-%. Normally officially published average statistics are based on weighted arithmetic averages and we think it is most natural choice also for new blocks of

flats and terraced house prices. So, our benchmark statistics is the weighted arithmetic average for which other aggregates are compared.

6.2 Decomposition of Actual Price Change for Stratum

The index number theory provides two main strategies for construction of index series: the base and the chain. Based on our multi period identity tests (Vartia, Suoperä, Nieminen & Montonen, 2018a) the chained type strategies almost always contain the chain error (or drift) that is contingent on data in question; somewhere the bias is harmless and somewhere severe. The base strategy is free of chain error, so we choose it as benchmark strategy for construction of index numbers. In the base strategy the base period may be defined various ways, e.g. certain week, month, quarter or for example previous year. We recommend in this study a previous year normed to average quarter as a base period. In practice this means that we are interested in direct price-links, that is $0 \rightarrow t$, where base period 0 is a previous year normalized as average quarter and period t is a quarter of a current year. Always when a first quarter of a current year appears, we change our base period to previous year normalized as average quarter.

Figure 7: Actual (*A*) and quality adjusted (*QA*) price changes (weighted arithmetic average) in 'Helsinki, two-rooms' 2010 = 1

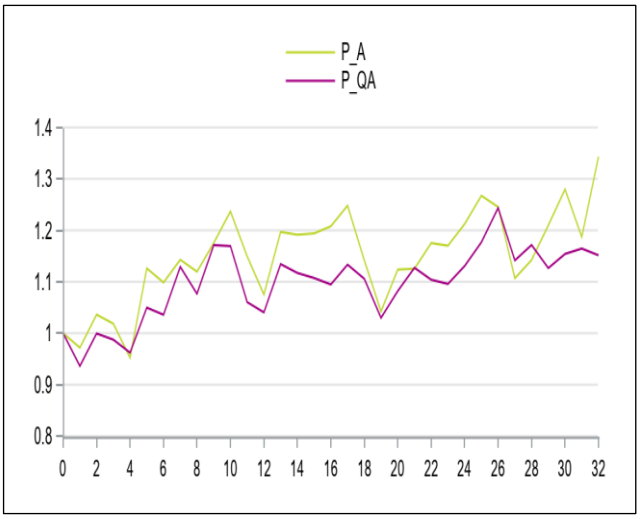


Figure 8: Corresponding indices for quality corrections (QC_{x_1}, QC_{x_3}, $QC_{x_5}QC_x = all\ together$) in 'Helsinki, two-rooms', 2010 = 1

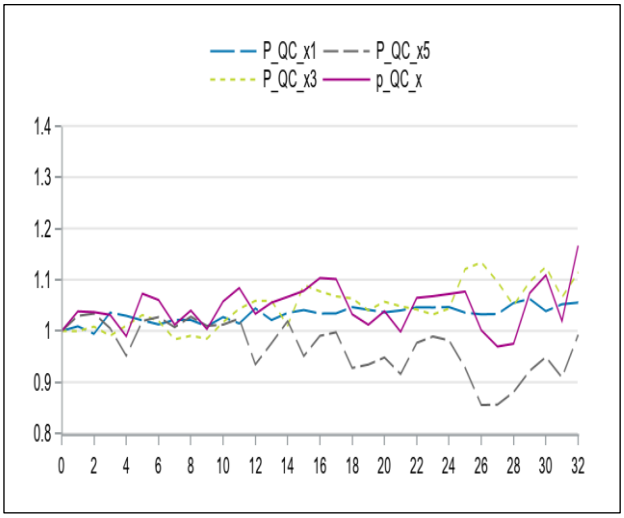


Figure 9: Actual (*A*) and quality adjusted (*QA*) price changes (weighted arithmetic average) in 'Espoo, two-rooms' 2010 = 1

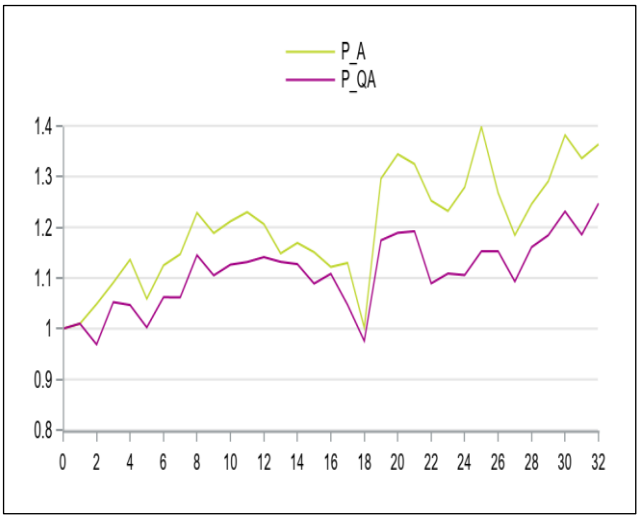
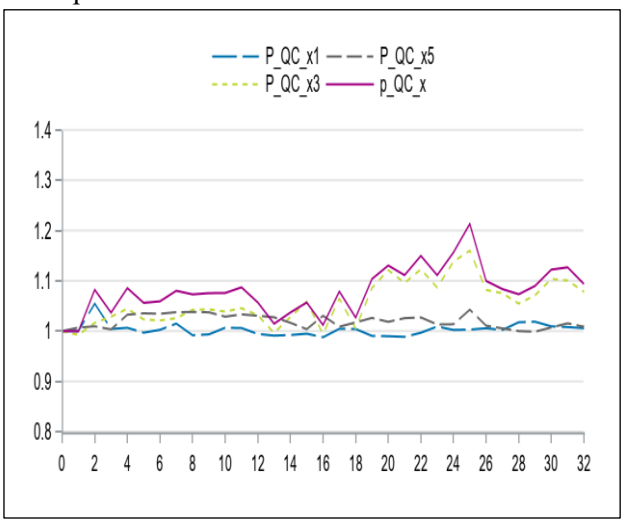


Figure 10: Corresponding indices for quality corrections (QC_{x_1}, QC_{x_3}, QC_{x_5}, $QC_x = all\ together$), in 'Espoo two-rooms' 2010 = 1



The decomposition (12) holds for every stratum. Practically this means that multiplying red lines for every *t* for example in Figures 7 and 8, we get the yellow lines in Figures 7 (i.e. $p_A^{t/0} = p_{QC,x}^{t/0} \cdot p_{QA}^{t/0}$, where $p_{QC,x}^{t/0} = p_{QC,x_1}^{t/0} \cdot p_{QC,x_3}^{t/0} \cdot p_{QC,x_5}^{t/0}$).

The flat size (square meter), distance and owner of a building lot have negative effect on prices – when values of $\bar{\boldsymbol{x}}^t$ variables exceeds (go under) their standard quality point $\bar{\boldsymbol{x}}^0$ we should correct actual prices upward (downward) in period *t* to quality adjusted level evaluated at standard quality point $\bar{\boldsymbol{x}}^0$. Both left-hand Figures shows that quality adjusted price changes are almost all the time below the price change of actual prices. This means that quality of flats has 'increased' especially for stratum 'Espoo two-rooms' from 2010 to 2018/4. Quality corrections behave heterogeneously for different stratums being significant role in determination of quality adjusted price changes.

6.3 Decomposition for Price Index Numbers

Index number formulas defined in Table 6 are used when strata decompositions are aggregated into crude aggregates - categories like 'One-room in Finland', 'Terraced Houses in Finland' etc. This will be done by following steps:

1. Take logarithm of (12) (additive form such that all price ratios in (12) are in logarithmic form).
2. Use index number formula (Table 6) in logarithmic form.
3. Calculate log price change for each component of (12) separately using same index number formula.
4. Take exp-transformation of each log price ratios.
5. Do the steps 1 to 4 separately for each average $m = uA, wA, uG, wG$.

Steps 1 – 5 gives us the hedonic price index number decomposition

(13) $$P_A^{t/0} \equiv P_{QC,x_1}^{t/0} \cdot P_{QC,x_3}^{t/0} \cdot P_{QC,x_5}^{t/0} \cdot P_{QA}^{t/0}$$

It is important to keep index number formula *P* and average *m* fixed for steps one to four.

Figure 11: Quality adjusted (*QA*) price index series (*m* =weighted arithmetic average) for One-room in Finland, 2010 = 1

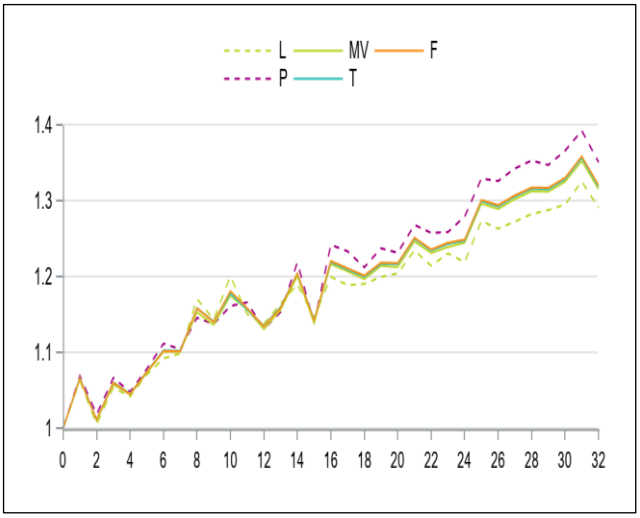
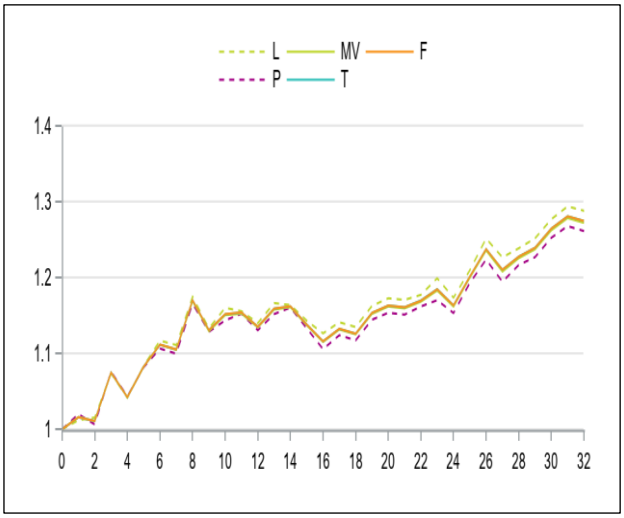


Figure 12: Quality adjusted (*QA*) price index series (*m* = weighted arithmetic average) for Three-rooms or more in Finland, 2010 = 1



First, we compare basic index numbers (*L* = Laspeyres and *P* = Paasche) to excellent ones (*MV* = Montgomery-Vartia, *T* = Törnqvist and *F* = Fisher). Figures 11 and 12 show that basic index number formulas are *data contingently biased* (see Vartia & Suoperä, 2018). The data contingent nature of bias is seen clearly - *L* is downward biased in left Figure but upward in right. These figures show that excellent index number formulas are very closely related. The same happens for any aggregation level. As result, the basic index number formulas should never be used, if excellent ones are available.

Because excellent index number formulas (see Vartia & Suoperä, 2018) are very closely related, any of them may be used. We demonstrate our results by Törnqvist formula. For 'One-room in Finland' the quality corrections are very significant – flats are smaller, distance from point of municipal services is further and building lot is not so often rented – so, index series constructed by actual price changes (yellow line) exceed significantly index serie constructed for prices being comparable in quality (red line). The quality corrections together increase up to 15 log-% (red line in Figure 14).

Figure 13: Actual (*A*) and quality adjusted (*QA*) price changes (weighted arithmetic average) 'One-room' in Finland, 2010 = 1, (P = Törnqvist).

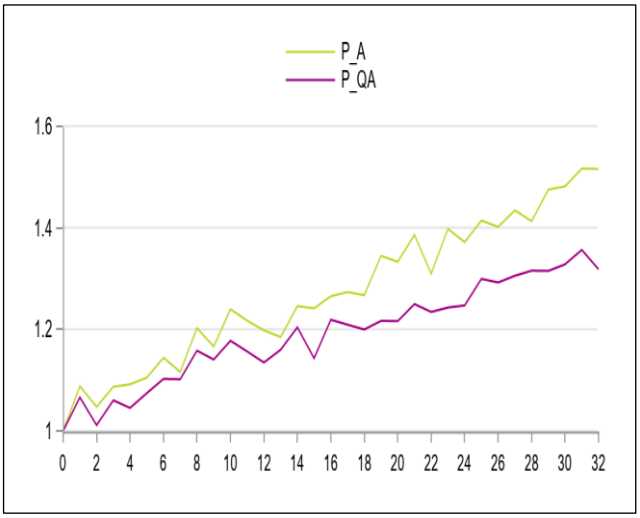


Figure 14: Corresponding indices for quality corrections (QC_{x_1}, QC_{x_3}, QC_{x_5}, $QC_x = all\ together$) 'One-room' in Finland, 2010 = 1, (P = Törnqvist).

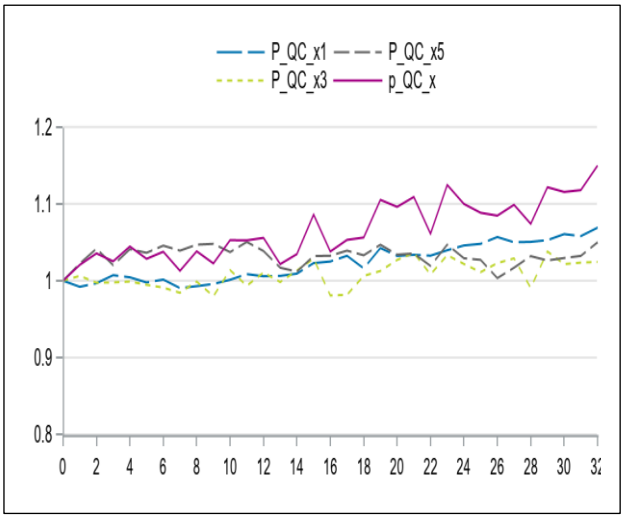


Figure 15: Actual (*A*) and quality adjusted (*QA*) price changes (weighted arithmetic average) 'Two-rooms' in Finland, 2010 = 1, (P = Törnqvist).

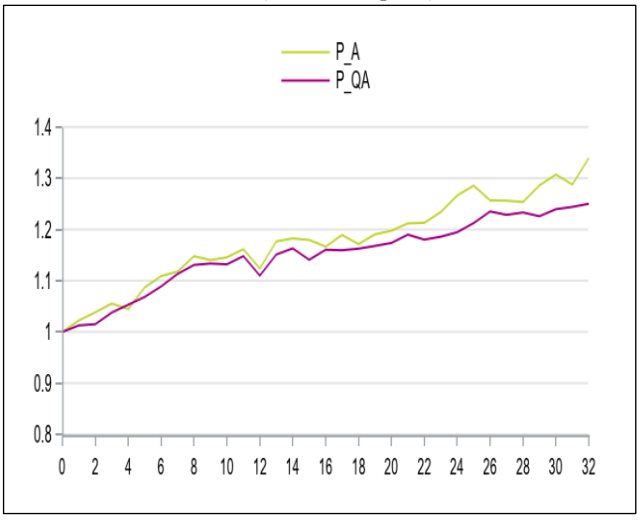
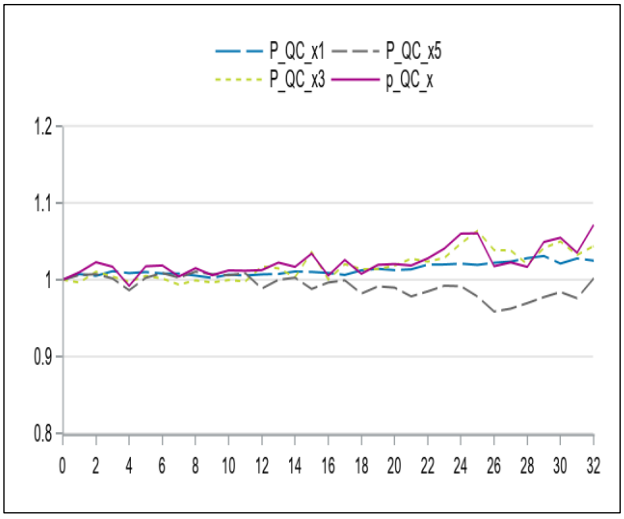


Figure 16: Corresponding indices for quality corrections (QC_{x_1}, QC_{x_3}, QC_{x_5}, $QC_x = all\ together$) 'Two-rooms' in Finland, 2010 = 1, (P = Törnqvist).



For ‘Two-rooms in Finland’ the quality corrections are not as significant role compared to ‘One-room in Finland’. Especially the share of rented lots starts to increase in middle of time span (grey dashed line in Figure 16).

Figure 17: Actual (*A*) and quality adjusted (*QA*) price changes (weighted arithmetic average) ‘Three-rooms or more’ in Finland, 2010 = 1, (P = Törnqvist).

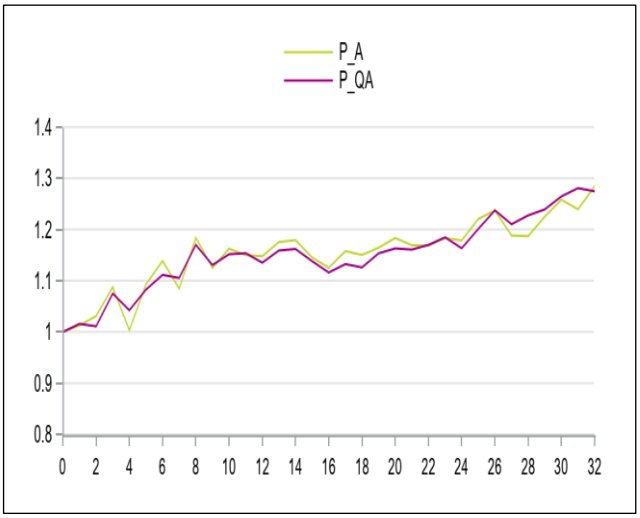
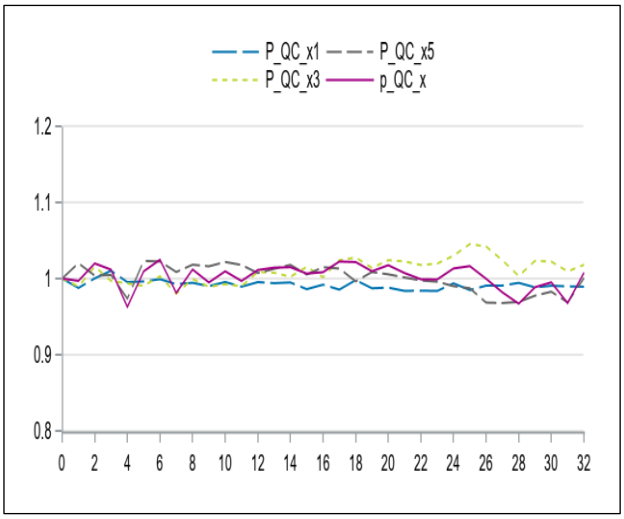


Figure 18: Corresponding indices for quality corrections ($QC_{x_1}, QC_{x_3}, QC_{x_5}, QC_x = all\ together$) ‘Three-rooms or more’ in Finland, 2010=1,(P= Törnqvist).



For ‘Three-rooms or more in Finland’ quality characteristics behave most stable and actual and quality adjusted price changes deviate not so much compared to ‘One-room in Finland’ and ‘Two-rooms in Finland’. The quality adjustment is necessary also for ‘Three-rooms or more in Finland’ because for some price-links quality characteristics deviate significantly.

The Figures 19 and 20 show how prices and quality characteristics have changed for ‘Terraced Houses in Finland’. The size of flat, distance and share of owner of a building lot increases in time. This means that actual prices in period *t* must increase to match the prices evaluated at standard quality point $\bar{\boldsymbol{x}}^0$. Contrary to ‘Block of Flats’, index series of adjusted prices exceeds index series constructed for actual prices.

Figure 19: Actual (*A*) and quality adjusted (*QA*) price changes (weighted arithmetic average) ‘Terraced Houses’ in Finland, 2010 = 1, (P = Törnqvist).

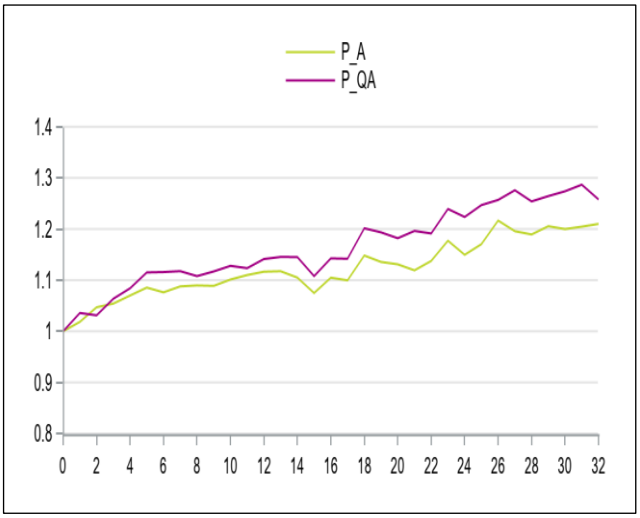
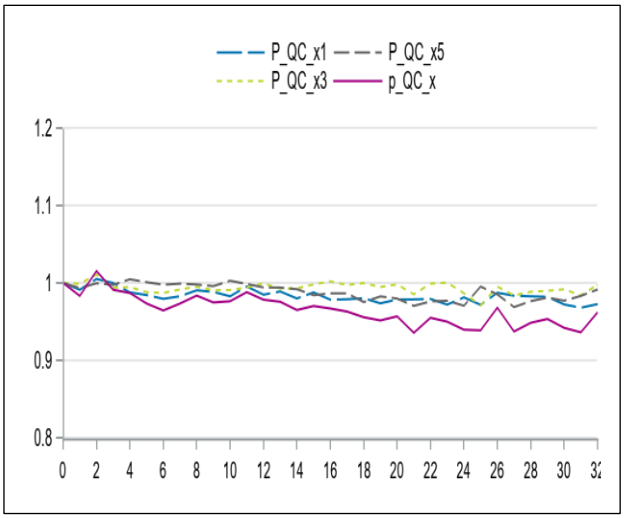


Figure 20: Corresponding indices for quality corrections ($QC_{x_1}, QC_{x_3}, QC_{x_5}, QC_x = all\ together$) ‘Terraced Houses’ in Finland, 2010 = 1, (P = Törnqvist).



7 Conclusion

Because transacted flat - new blocks of flats and terraced houses – emerges only once in data, index number theory based on bilateral or multilateral methods is not available for observations (i.e. matched pairs). Observations should aggregate into some appropriate partition, which in this study is based on 12 region each divided into four flat-type – block of flats into one-room, two-rooms and three-rooms or more and terraced houses into one class. Our partition includes 48 strata. We apply regression analysis to our data having described partition. Our regression analysis is based on heterogeneously behaving cross sections. We average our price models into strata level for each time period separately and define difference simply saying between periods 0 and *t*. Whatever average of prices we use, they are not comparable in quality for price-link $0 \rightarrow t$ – quality differences should be removed from actual or true price changes to get quality adjusted price change evaluated in standardized point of quality. For that we use so called hedonic or Oaxaca-decomposition.

Normally decomposition is based on standard textbook solution, where price ratio of unweighted geometric average prices is decomposed into *price change due to quality difference* and quality adjusted price change evaluated at *standard point of quality*. This solution is not satisfactory for officially published average statistics, because unweighted geometric average prices deviates from weighted arithmetic average prices sometimes more than 10 log-%. So, we develop two new theorem of price aggregation for semi-logarithmic price models. We apply these theorems for estimated OLS solution and get two new hedonic decompositions based on weighted and unweighted average prices. We show that similar decomposition is possible also for weighted geometric average prices. We derive hedonic price decomposition for four averages one of which is standard textbook solution for unweighted geometric average prices.

In strata level no index number formula is needed – hedonic price decomposition divides average price change into quality corrections and quality adjusted price change for all averages separately. We show that quality adjustment is necessary. Our benchmark decomposition is weighted arithmetic average.

We use index numbers first time, when strata level decompositions are aggregated into crude aggregation level. We use basic and excellent index number formulas. Our construction strategy of index series is base strategy, where base period is previous year normalized as average quarter. Index series constructed by this strategy is free of chain error. As conclusion, basic index number formulas are data contingently biased and should not be used as official statistic. Instead, excellent index number formulas are very closely related – any of them may be selected to official production.

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.

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

Nieminen, K. and Montonen, S, "The foundation of index calculation", (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html), 2018.

Koev, E. 'Constructing a Hedonic Wage Index: Pilot Study for the Finnish Metal Industry', EU Projects, 1997.

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.

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

Suoperä, A. 'Kirkon ja muiden sektoreiden suhteelliset palkkaerot: Niiden analyysi ja tilastointi', 2002. (study subject to a charge, Church labour markets/Statistics Finland).

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

Suoperä, A. 'Naisten ja miesten palkkaerot valtiosektorilla ja yleisillä työmarkkinoilla', 2010a (study subject to a charge, Ministry of Finance/Statistics Finland).

Suoperä, A. 'Kunnan ja yleisten työmarkkinoiden palkkaerot: Hedoninen menetelmä', 2010b (study subject to a charge, Municipal labour markets/Statistics Finland).

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

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 topics no. 25,1979.

Vartia, Y. 'Good choices in index number production', 2018 (http://www.stat.fi/static/media/uploads/meta_en/menetelmakehit-ystyo/over_identification_new_vartia2018.pdf).

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., Suoperä, A., Nieminen, K. & Montonen, S. "Circular Error in Price Index Numbers Based on Scanner Data. Preliminary Interpretations", 2018a. (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html)

Vartia, Y., Suoperä, A., Nieminen, K. & Montonen, S. "The Algebra of GEKS and Its Chain Error", 2018b. (http://www.stat.fi/meta/menetelmakehitystyo/index_en.html)

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. (in Swedish)

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

Appendix 1. Analysis and synthesis of price determination in Heterogeneous Cross-sections.

Analogously with Suoperä & Vartia (2011, p.11-18) we define the estimated regression models as

(1) $$\log(p_{irt}) = \hat{\alpha}_{rkt} + \boldsymbol{x}'_{irt}\widehat{\boldsymbol{\beta}}_{rt} + \hat{\varepsilon}_{irt}$$

where $\hat{\alpha}_{rkt}$ is estimated price effect for the flat type (i.e. k = 1, 2 ,3, 4) in the region r in time period t. The estimates of parameters $\hat{\alpha}_{rkt}$ and $\widehat{\boldsymbol{\beta}}_{rt}$ in (1) are OLS estimates used in analysis of hedonic price index numbers. All elements of the above equation are known for observation i and for period t – log-prices, quantities of quality characteristics, parameters and even residuals! We continue our analysis in spirit of Suoperä & Vartia (2011) by averaging over all observation level equations using the basic lemma of aggregation (Vartia, 1979, 2008) to get macro model, which we break back into observation level. We call such a method as being a solution backward. So, we get more operational representation for (1), that is

(2) $$\log(p_{irt}) = \hat{\alpha}_t + \boldsymbol{x}'_{irt}\widehat{\boldsymbol{\beta}}_t + (\hat{\alpha}_{rkt} - \hat{\alpha}_t) + \boldsymbol{x}'_{irt}\left(\widehat{\boldsymbol{\beta}}_{rt} - \widehat{\boldsymbol{\beta}}_t\right) + \hat{\varepsilon}_{irt}$$

The new representation of equation (5) decomposes the regression model into two parts: A representative data generating process for all transacted flats $\hat{\alpha}_t + \mathbf{x}'_{irt}\widehat{\boldsymbol{\beta}}_t$ and two terms describing observation specific behavior as deviation of the representative one $(\hat{\alpha}_{rkt} - \hat{\alpha}_t)$ and $\mathbf{x}'_{irt}\left(\widehat{\boldsymbol{\beta}}_{rt} - \widehat{\boldsymbol{\beta}}_t\right)$. 'The equation (2) is a reparametrized version of (1) – only their arguments are decomposed differently. In the former the heterogeneity is distributed among all observation units and in the latter this is separated into its own terms. The price equation for transacted flats consists of two sets of variables: The first set includes the exogenous independent variables $(1; \mathbf{x}'_{irt})$ and the other all the 'covariates' $(\hat{\alpha}_{rkt} - \hat{\alpha}_t)$, $\mathbf{x}'_{irt}\left(\widehat{\boldsymbol{\beta}}_{rt} - \widehat{\boldsymbol{\beta}}_t\right)$, which are the microelements of the covariance terms distributed element by element in the observation level. Dimensions for the both sets of variables are K+1, where K is the number of exogenous independent variables. Next, we free all the parameters of (2), including the unities of the heterogeneity terms, and form the second stage estimation equation;

(3) $$\log(p_{irt}) = \hat{\alpha}_t + \mathbf{x}'_{irt}\widehat{\boldsymbol{\beta}}_t + (\hat{\alpha}_{rkt} - \hat{\alpha}_t)\cdot\gamma_t + \mathbf{x}'_{irt}\left(\widehat{\boldsymbol{\beta}}_{rt} - \widehat{\boldsymbol{\beta}}_t\right)\cdot\lambda_t + \hat{\varepsilon}_{irt}$$

or equally by vector and matrix notation

(4) $$\mathbf{y}_t = \mathbf{X}_{1t}\boldsymbol{\beta}_{1t} + \mathbf{X}_{2t}\boldsymbol{\beta}_{2t} + \hat{\boldsymbol{\varepsilon}}_t$$

The first column of matrix $\mathbf{X}_{1t}$ is a vector of ones (i.e. constant) and the other columns are correspondingly micro explanatory variables (or quality characteristics) of the price model. The variables of the matrix $\mathbf{X}_{2t}$ are the covariates $(\hat{\alpha}_{rkt} - \hat{\alpha}_t)$, $\mathbf{x}'_{irt}\left(\widehat{\boldsymbol{\beta}}_{rt} - \widehat{\boldsymbol{\beta}}_t\right)$. The two estimates of the (K+1) dimensional vectors of parameters are $\boldsymbol{\beta}_{1t} = \left(\hat{\alpha}_t; \widehat{\boldsymbol{\beta}}_t\right)$ and $\boldsymbol{\beta}_{2t} = (\gamma_t; \boldsymbol{\lambda}_t) = (1; \mathbf{1})$ vector of ones. The i:th observation in the equation (4) is exactly the i:th observation in the equation (2). For example, the i:th element of the residual vector $\hat{\boldsymbol{\varepsilon}}_t$ is exactly the OLS i:th residual estimated in the analysis stage (i.e. in (1)). So, the equation (4) is based purely on 'bookkeeping', because it is formed of known equations and their estimated parameters and is rewritten in the formulae (2,3 and 4). In the equation the first term on the right, $\mathbf{X}_{1t}\boldsymbol{\beta}_{1t}$, indicates the representative behavior of all observations, while the term $\mathbf{X}_{2t}\boldsymbol{\beta}_{2t}$ contains observation by observation the heterogeneous behavior differing from the representative agent' (Suoperä & Vartia, 2011, p.14-15).

The equation (3) or (4) is just a rewritten original regression model including 12 (r =1,…, 12) separately estimated regional price models (1). In addition to duplicating previous parameter estimates in (4) compared to single separate regional price model, we get standard errors for average macro parameters $\left(\hat{\alpha}_t; \widehat{\boldsymbol{\beta}}_t\right)$ – *that is at last a new result?* It may come first as a surprise that this OLS-model (4) must exactly replicate all the previous average parameter values and give the unity coefficients for the covariates. The reason for this is purely algebraic in character. We have in the this OLS estimation (4 all the sufficient information to produce OLS-solution not only for the combined regions, but also for all its separate ones. Because (4) is capable of producing the previous OLS-solution (1) with its overall minimum sum of squares, this actually is its OLS-solution. All

other parameter estimates would give a larger sum of squares. The OLS-estimation (4) replicates in this way all the previous region wise regressions of the analysis stage: even the residuals are identical in them.