Residential Property Price Index

RPPI

Practical Compilation Guide

2020

CONTENTS

TABLES

ACKNOWLEDGEMENTS

INTRODUCTION

1. The *Guide* sets out practical advice on the compilation of a Residential Property Price Index (RPPI) based on the conceptual approach described in the *Handbook on Residential Property Price Indices* (*Handbook*), published in 2013.

2. The *Handbook* notes that "*Residential property price indices have a number of important uses:*

- *as a macro-economic indicator of economic growth;*
- *for use in monetary policy and inflation targeting;*
- *as an input into estimating the value of housing as a component of wealth;*
- *as a financial stability or soundness indicator to measure risk exposure;*
- *as a deflator in the national accounts;*
- *as an input into an individual citizen's decision making on whether to buy (or sell) a residential property;*
- *as an input into the consumer price index, which in turn is used for wage bargaining and indexation purposes;*
- *for use in making inter-area and international comparisons.*"

3. The Second Phase of the G-20 Data Gaps Initiative and guidance on Financial Soundness Indicators identify real estate statistics, in particular price changes on residential property, as a critical input into financial stability policy analysis and macroprudential measures. Data users in IMF member countries have expressed a need for RPPIs and the IMF has provided training and capacity building to many member countries.

4. The *Guide* was prepared by the IMF Statistics Department, building on experience gathered during technical assistance and training missions, delivered to RPPI compilers from more than 80 countries over the past six years. Compilers have expressed the need for practical guidance to complement the theoretical content of the *Handbook*, particularly when transitioning from less to more sophisticated methods.

5. The *Guide* is written for data compilers in statistical offices, central banks, housing agencies, or other statistical agencies and demonstrates how to compile and disseminate RPPIs. Using the *Guide* requires a background in statistics and some knowledge of econometrics.

6. The *Guide* includes several practical exercises compilers can complete to improve their understanding of the data sources and methods used to compile an RPPI. The exercises are delivered through a series of step-by-step instructions that make use of a synthetic data set built by the Central Statistics Office Ireland (CSO) adapted by the IMF and R scripts (see box below) developed specifically for the RPPI compilation.

7. The exercises accompanying this Guide utilize a synthetic dataset generously shared and developed by the CSO Ireland. The Synthetic House Price Dataset was generated by CSO Ireland

by applying several statistical masking techniques to create a dataset which is anonymous and materially different from the Irish RPPI raw data, but which preserves the aggregate correlations between the variables which affect house prices. The techniques used in constructing the synthetic data prevent the identification of individual transactions or dwellings. As the Synthetic House Price Dataset is materially different to the Irish RPPI dataset, it cannot be used to replicate the Irish RPPI results.

8.   The IMF will provide the Synthetic House Price Dataset, upon request to compilers, researchers and students who agree to the following conditions of use.

- The Synthetic House Price Dataset will only be used for training and educational purposes.
- The CSO Ireland is to be acknowledged as the source of the Synthetic House Price Dataset.
- In all coursework material, training workshops and related reports the dataset is referred to as the Synthetic House Price Dataset, or simply as a synthetic dataset, not as Ireland's Residential Property Price Index dataset.
- Acknowledgement that the CSO Ireland is not liable in any way for the usage of the Synthetic House Price Dataset. All analyses, inferences and interpretations of the data and all consequences thereof are entirely the responsibility of the data user.
- Neither the Synthetic House Price Dataset, full or partial copies of the dataset, are to be made available to any third party.

9.   Please email a request for access to the synthetic dataset to the Division Chief of the Real Sector Division of the Statistics Department (rppi@imf.org) of the IMF. The request should include information on the reason for access (compilation, research, study), organization details (if any) and other relevant information. A register of users who receive the Synthetic House Price Dataset will be maintained by the IMF.

Box 1. R Software

R is a programming language and free open-source (R-Studio) software environment for statistical uses, with a large online community that supports its availability and reliability and provides technical support. The hardware and software requirements are the same as for other econometric software such as STATA, SAS or E-views. The R software is open which means that any person can use it free of charge. The R scripts accompanying this *Guide* have been prepared such that compilers should be able to use them with their own data to compile a monthly or quarterly index, with a few and easy-to-perform adaptations.[2] Using the R program and the data set can help compilers to develop their capacity for index compilation.

---

[2] Specific guidance on how to make these adaptations is provided in each R script.

10.　　The *Guide* is useful for all RPPI compilers but specifically targets organizations with capacity constraints (lack of econometricians, programmers, or IT support) compiling new RPPIs. The aim of the *Guide* is to effectively help developing countries close data gaps and improve data quality. Countries that are already compiling an RPPI can also benefit from the *Guide* by using it to improve their methodologies or processes.

11.　　The *Guide* is divided into three sections. The first section illustrates the type of data commonly used to construct RPPIs and how compilers can best evaluate the quality of the source data. It also outlines how to prepare or pre-process data prior to compiling an RPPI. Finally, the section stresses the importance of selecting appropriate weights and the different type of data that can be used to construct the weights.

12.　　The second section of this *Guide* outlines the different methods that can be used to compile RPPIs.

13.　　Finally, the third section provides guidance on dissemination practices and suggests some additional indicators compilers may want to construct to complement the RPPI.

DATA SOURCES AND WEIGHTS

A. Identifying Data Sources

Reference: Handbook on Residential Property Price Indices, pages 102-111

14.   When initially designing an RPPI program, compilers should target a national monthly or quarterly index covering monetary transactions (mortgage and cash purchases) of all types of residential properties (single and multi-family dwellings, new and existing dwellings) in line with the concepts outlined in the *Handbook*. In cases where the available data, operation of the market or other constraints prevent complete alignment with the *Handbook*, compilers may need to reduce the coverage and/or frequency of the index. For example, an index that only covers urban markets or registered transactions may be considered in these cases.

15.   To compile an RPPI conforming to the recommended best practices data related to the transaction price as well as key characteristics[3] of the dwellings are required to assure a constant-quality index. As with any price index the target is to follow the price trend by comparing "like-with-like". This is a challenge since a direct price comparison requires comparable dwellings or the same dwelling to be transacted in consecutive periods. This is generally not the case with residential properties where the same residential property may only be sold once every 20 years. Given their infrequent sale and the heterogeneity of residential properties, quality adjustment techniques are required to derive measures of pure price change. This means that the data requirements for a high quality RPPI are extensive and rely heavily on the acquisition of detailed information about each property given that there are a wide range of characteristics that can influence the price of a dwelling.

16.   The most commonly available characteristics are the dwelling type, floor area, number of rooms, district, region, property type, age, vintage (new or existing), completion state of dwelling, existence of an elevator, etc. Each of these characteristics may affect the price of a dwelling.

17.   Location is the most important characteristic and can be described with categorical variables for province (state), city and sub-city areas. It is desirable to include more granular areas when possible, however it is often a challenge to acquire data at this level of detail. There are two potential limitations when granular location data are not available. First, given that high- and low-value areas within the city are lumped into one category, there could be systematic variation by sub-area over time that is not being captured, so if low-value areas trade more over time, this may be "measured" as a decline in prices. Second, different sub-areas may have different price trends (as well as different price levels) that are being averaged into one trend when considered together. Other potential location-specific factors could be included in an

---

[3] Key characteristics refer to structure and location attributes such as usable floor area, number of rooms, district, region, property type, age, new or existing, completion state, renovations, etc.

ideal/advanced analysis (such as access to transport facilities or other amenities).

18.   Dwelling relevant characteristics may vary substantially by market. For example, heating is a key variable in northern European countries but irrelevant in Africa. Size appears to be important in all markets but how size is measured may vary and can be multi-dimensional, e.g. number of rooms, floor area and site/lot size for houses.

19.   As a first step in designing an RPPI, compilers should acquire an in-depth understanding of the real estate transaction process within their country. This will allow the compiler to understand the type of information generated throughout the transaction process and determine which data are useful for the purposes of compiling an RPPI. Figure 1 illustrates a possible set of information that could be generated leading up to a real estate transaction, during the execution of the transaction, and after the transaction has occurred.

Figure 1. Example of a Timeline of the Transaction Process
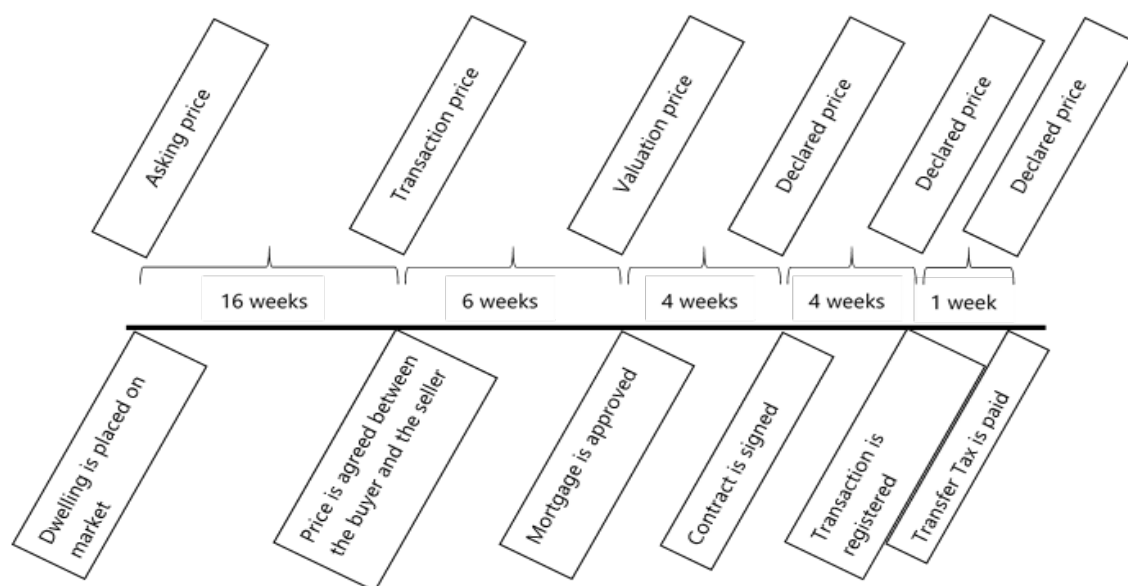
20.   It is clear from Figure 1 that the same dwelling can have many different prices at distinct points in time. For example, there is typically an asking price, a transaction price, and a declared price. When constructing an RPPI the compiler should ideally aim to acquire the agreed upon price that corresponds to the date the transaction is finalized. This is referred to as the transaction price.

Box 2 Price Concepts and Their Fitness for Purpose

1. Asking price – Corresponds to the price asked when the dwelling is put on the market. It may change if the dwelling takes some time to be sold and after negotiation. In some markets, the asking price can be higher than the target price, although as the purpose of the RPPI is to capture changes, rather than levels, even in these circumstances, it may be a suitable proxy see Lyons (2019) for more on this.

2. Transaction price – After negotiations the buyer and the seller agree on the price. This is the target price, e.g., the price actually paid. However, this can be difficult to obtain since at this moment there is no register in a formal or institutional database.

3. Valuation price – The dwelling is evaluated by the lending institution. The value of the appraisal can be higher, lower or equal to the transaction price. If the value of appraisal equals the value of the transaction, these data are suitable, but otherwise are not recommended.

4. Declared price – This is the "legal" price of the dwelling, normally used for tax purposes. The registration of the selling price will normally happen later than the transaction date and declared prices may differ from the transaction prices. Depending on the size of these differences it can be or not a suitable proxy.

---

21. The data required to construct an RPPI may be obtained from a variety of sources including websites, real estate authorities or surveys conducted by NSOs. The most commonly used data sources are illustrated in Figure 2 along with the price concepts most commonly associated with the data source.

Figure 2. Data Sources and Price Concepts
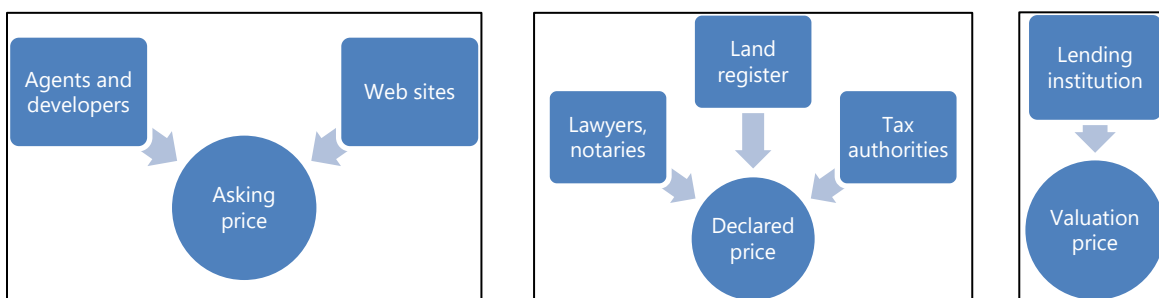
22. In Figure 2, the target price, i.e. the transaction price, is likely to be the declared price or may be obtained also from the lending institutions.

23. Data from websites can be obtained by web scraping or through direct provision by the site owner. Additionally, data can also be obtained via application programming interfaces (APIs) when the website or data owner decides to make the data available via an API. In general, websites provide a wide coverage, a detailed description of the property, and the asking price. In many countries, online sites cover the entire country, all types of dwellings and both cash and mortgage transactions. Also, in many cases dwellings sold through agents, developers, or owners, are included. Timeliness is an advantage with this source since the price is obtained as soon as the property enters the market.

24.   While data obtained from websites have many advantages there are often several drawbacks. In some countries, web listings are confined to higher-end properties that are only transacted by developers. It is also common that the asking price differs from the final transaction price which is not captured in the advertisement. However, the weight of academic research finds that changes in quality-adjusted asking prices, are very accurate proxies for trends in the underlying transaction prices. Nevertheless, price indices based on asking prices may fail in anticipating turning points in the housing market. Finally, compilers cannot always assume a property has sold if it is no longer listed. A dwelling can be "delisted" simply because the owner decides no longer to sell.

25.   Compilers can also survey agents and developers to obtain residential real estate prices and related information. The advantage of a survey is the compiler can design the questionnaire to obtain the precise information they require according to the concept they are measuring. While a survey provides the compiler with the greatest degree of control over the information, they use to construct the index, it has several drawbacks including high costs and response burden. In addition, it is very difficult to obtain representative samples as explained in Box 3.

Box 3. Sample Representativeness

Sample surveys can be used to collect data from agents, developers, lawyers, or notaries. This raises the issues of sampling bias and representativeness. When sampling respondents, care must be taken to guarantee an adequate coverage of the turnover, geography and dwelling type. The sample should be stratified with the larger developers and agents included in the sample on a regular basis and the smaller agents and developers rotating into the sample every year. Information on turnover may be available from business surveys or government administrative records. The larger agents and developers may work exclusively with dwellings located in the main cities. To ensure adequate coverage of the entire territory small developers and agents may need to be included in the sample. In addition, small developers and agents may focus on a niche part of the market that may be missed if they are not included. Finally, surveys can be costly and incur respondent burden. More information on sampling can be found in Chapter 4 of the *CPI Manual: Concepts and Methods,*

---

26.   A key data source that compilers can use to compile RPPIs are national land registries or property assessment files maintained by government or tax authorities. These data may cover all transactions in all geographic regions. Depending on national circumstances, the quality of the data may be good given their coverage and the fact that they are used for legal and tax purposes. Administrative data is also often the least costly option and does not place a significant burden on respondents. While land registries and property assessment files are often an excellent source of information, care should be taken to ensure the transactions are recorded in the proper period. In certain countries it is possible that the registration of the sale may take place three months or more after the transaction takes place. In this case the price would be recorded in the wrong period and the compiler may introduce a time bias in the calculation of the index. The definitions, classifications and underlying concepts associated with register data may also differ from those required for statistical purposes. It is therefore essential that compilers obtain access to metadata describing the source, concepts and definitions used by the registry.

When the concepts underpinning the registry differ from those needed for the RPPI, some adjustments (or filtering) of the data may be necessary. Another issue is that registers usually do not contain detailed information about the characteristics of the property which is needed when using hedonic methods.

27.   Another possible source of information to compile RPPIs is mortgage data. These data are often timely, and the geographic coverage is generally broad. However, cash transactions are excluded. Lending institutions often hold a large amount of information related to the characteristics of the dwellings which can also aid in the calculation of the RPPI. In less developed countries, the mortgage sector may have very weak coverage – especially outside the major city/cities.

28.   When using mortgage applications as a data source for compiling RPPIs, compilers should ensure that the price on the application reflects the final selling price. Mortgage applications often contain the upper limit of the value that can be borrowed which can be different from the agreed upon transaction price of the property. Additionally, in some countries the banking sector is very fragmented, and a substantial amount of work will need to be undertaken to standardize the data so that it can be processed into an RPPI. In some countries it may be best to work with the Central Bank who may act as a clearinghouse for detailed mortgage information.

29.   It may also be possible to obtain information from lawyers or notaries. Contracts are often registered with lawyers or notaries following the sale of a residential property. These contracts are typically good data sources in terms of coverage and contain detailed information about the location and the actual transaction price of the dwelling.

30.   However, if sales are registered with a lawyer, it may be difficult to obtain a complete electronic file that includes all transactions for a given period. This is because in some countries lawyers are not required to register their contracts into a national database. When a national registration system does not exist, each lawyer may have their own contract registration system. The compiler would then need to obtain the registry from each lawyer and integrate the data into a common registry which is costly and inefficient. Obtaining data from lawyers may also be difficult as lawyers in most jurisdictions cannot provide information about their clients to authorities. Notaries are normally regulated and required to record the sale of a property in a common database. The contract is likely to be the first official register that secures the transaction and the transfer of ownership of the dwelling. In most cases the price that is recorded in the database is the target price for the RPPI. The main drawback with this data source is that it may often lack detailed information on the characteristics of the dwelling.

31.   The compiler should always bear in mind that the available data sources and their strengths and weaknesses are country dependent.

Table 1. Comparison of the Features of the Different Data Sources

Data Provider	Availability	Coverage	Timeliness	Detail for Quality Adjustment	Limitations of Observed Price
Websites	Good. Immediate and free.	Partial	Best (Within the day)	Must be checked on each website	Asking prices often differ from transaction prices
Agents	Good. Requires a survey therefore is costly.	Partial	Good	Good	Asking prices often differ from transaction prices
Developers	Good. Requires a survey therefore is costly.	Partial	Good	Good	Asking prices often differ from transaction prices
Lending institutions	Must be negotiated	Excludes cash transactions	Good	Good	Value can differ from transaction price
Lawyers or notaries with electronic data set	Good	Good if mandatory or very common	Good	Must be checked	May suffer from underreporting of values
Lawyers or notaries without electronic data set	Costly	Sample	Good	Must be checked	May suffer from underreporting of values
Land register and Tax authorities	Good	Complete	Can be rather late	Must be checked	May suffer from underreporting of values

B. Assessing Source Data Quality

32. Prior to constructing an RPPI and selecting the data sources, compilers should undertake a thorough assessment of the quality of the data. It may be useful for the compiler to assess the data source using a quality assessment framework. Statistical authorities often have quality assessment frameworks that are used to assess statistical programs within their jurisdiction. If one exists compilers should use their country-specific quality assessment framework to assess the quality of their source data. If a data quality assessment framework does not exist the compiler should, at a minimum, assess the data in terms of *accuracy*, *fit for purpose, structure*, *coverage, and availability*.

33. The *accuracy* can be determined by assessing the occurrence of errors and missing observations.

34. *Fit for purpose* is related to conceptual issues such as whether the price reflect the target price concept or if the coverage is adequate (national vs local) or if it meets the desired timeliness and frequency (monthly, quarterly, annually) requirements.

35. The ***structure of the data*** can be measured by the number of variables or attributes associated with each observation. For example, source data only containing price and city is less useful than source data containing, price, exact location within city, size, type of structure and age.

36. The ***coverage*** examines the share of observations included in the source data as a share of the total observations required (or possible). For example, if a compiler is targeting a national index but the data source only includes transactions for two of the 10 largest cities then the coverage would be low (even if they had a census of transactions for the two cities). In a highly regulated market, the coverage should be high. Conversely, in a less regulated or unregulated market the transaction can be as simple as the two parties agreeing on a price and the corresponding amount changing hands. It is highly likely that in a regulated system at least some components of the transaction will be registered, and the coverage will be high. In an unregulated market it is less likely that any component of a transaction will be registered, and the coverage may be low.

37. Compilers also need to assess the **availability** of the data. While data may be highly accurate with good coverage, if the data cannot be acquired in a reliable fashion according to an agreed upon schedule it will be of little use. Additionally, data quality is lower if records are in a format that the compiler cannot readily access. For example, in some cases the information may only be available in paper format. When the transactions are registered on paper the information must be digitized. In other cases, transactions may be recorded in digital form, but several different regional systems may be in use, thus several data files using different standards may need to be standardized and merged into a unique data file. Examples of specific challenges include:

- dealing with different formatting;

- varied variables;

- variables with the same name but diverse meanings; and

- lack of a common key to integrate the information.

38. Once the RPPI compiler has assessed the source data and determined if it can be used to develop an RPPI, he/she should enter into a formal data exchange agreement with the data provider to ensure a consistent and reliable transmission of the source data. This agreement should articulate, among other things:

- procedures for the secure transmission of the data (since the data often include sensitive information);

- the provision of metadata;

- an agreed upon schedule for the transmission; and

- validation reports to ensure the data were transmitted correctly.

39. The RPPI compiler should also ensure that a data management system is in place to store, process, and back up the files. The compiler should keep track of the files that were used when calculating the RPPI to guarantee the replicability of the index compilation. The storage space required for all the files can be quite large. It is recommended that the compiler maintain and archive the raw data, the cleaned data, and the processed data for each compilation cycle. If a proper data management system and practice does not exist, the data can be damaged or even lost.

40. Source data used to construct RPPIs are often sensitive. The compiler's ability to access the micro-data may be limited even when regulations are in place that entitle statistical offices to have access to data. For instance, tax authorities may be reluctant to share data with other institutions. In these cases, the compiler should work with the data provider to address confidentiality issues. It is considered a best practice to establish a memorandum of understanding that governs the exchange of the information and clearly lays out each party's responsibility with respect to protecting the confidentiality of the data.

41. It is recommended that the compiler develop standard validation reports that can be generated upon receipt of the data. In addition to developing validation reports the compiler may also want to open the file for a brief visual inspection. This will allow the compiler to detect severe errors such as:

- an empty or corrupted file;

- incorrect time scope of the data set;

- an unexpected number of observations;

- a file with unrecognizable characters;

- binary variables with numbers other than zero or one; or

- to confirm that the structure of the data remains stable.

42. The prices should usually be recorded in the official local currency since this is usually the actual price facing households. If the original data file shows other currencies use the daily exchange rates from the national central bank website to convert the transaction value into the local currency.

Box 4. Preparing Data for Processing

Before importing a data file (excel or csv) into a processing system some preparations should be undertaken to make sure the program reads the file correctly. The required preparations are specific to the software and the dataset being used; however, some general advice and examples are provided.

- Text in the data file, such as names of variables (e.g., floor area) and occurrences of qualitative variables,[4] should not have spaces, accents, and misleading characters like minuses (-). The use of underscores is helpful - "floor_area" and "semi_detached" since it facilitates readability by machines and humans.
- Numbers must be formatted as "number" and qualitative variables as "text."
- Commas or dots to split the thousands should not be used.
- The names of the variables must be the same in every file. For example, the software will read "detached_house" as different from "Detached_house" (Capitalization) and "Floor_Area" as different from "FloorArea" (missing the underscore).
- It is advisable to replace all missing values with "NA" since the software often does not read empty cells correctly.
- Normally the data set includes a column with the date to which the transaction refers. Adding a column with the identification of the quarter (for quarterly indices) or the month (for monthly indices) is recommended since it is often the case that the data file of the following quarter will include transactions that happened in the previous quarter.

Exercise 1: Assessing Source Data Quality in Practice

43.   The following exercise outlines a series of assessments that a compiler may undertake to analyze the quality of a data source and pre-process their data so that they are ready to be used to develop an RPPI. These steps are examples and are by no means exhaustive. The precise steps that a compiler includes will depend on the data sources and methods they are using to develop their index as well as their specific institutional arrangements.

44.   The R script "data_analysis.r" and the data set "1Q2008_raw.csv" that accompany this *Guide* can be used to complete this exercise. The script file includes detailed comments.

45.   **Visually inspect the data** – Table 2 lists the variables that are included in the synthetic set of data. As a first step the user should open the data set and visually inspect the data and identify any problem cases that may require further investigation.

---

[4] A qualitative variable, also called a categorical variable, is a variable that is not numerical. It describes data that fit into categories. (e.g., semi-detached). A numeric variable can also be treated as categorical when its values are grouped in categories.

Table 2. Data Structure

Data structure	Description	Type
Period	Month, quarter, and year	Categorical
Id	Unique identifier of the dwelling	Numerical
Year	Year	Categorical
Month	Month	Categorical
Dwelling_Type	Type of property	Qualitative/Categorical
Price	Registered price of the dwelling	Numerical
Floor Area	Total floor square meters (including common areas)	Numerical
Neighborhood_Affluence	*See text just below this table**	Numerical
Year Built	Interval within the dwelling was built	Categorical
BER	Building Energy Rating	Categorical
County	County where the dwelling is located	Qualitative/Categorical
Region	Region where the dwelling is located	Qualitative/Categorical
Status	New/Existing	Categorical
Central_Heating	Heating system existing in the dwelling	Qualitative/Categorical
Neighborhood_Type	Rural / Urban	Categorical
Building_Levels	Number of story's in the dwelling	Categorical

*The variable Neighborhood Affluence (NA) is a synthetic index. The Index refers to a small area where the house is located, not to the house itself. Small areas are the smallest geographical units comprising 80 to 120 households. The full methodology is quite complex, but in short, it considers the demographic profile, social class composition and employment in each of the small areas, as measured by a census of population.

46.   Check the size of the file – The size, the variables and the number of observations in each file should be compared with the previous files to assess the reasonability of these indicators. The variables should be always the same.

47.   Analyze summary statistics – Compilers should assess the economic reasonability of the data by creating summary statistics for all variables in their data set. The tables below provide an example of the type of summary statistics, derived from the synthetic dataset, that can be calculated. From Table 3 we observe that there is a large range between the minimum and the maximum price. This is a good indication that the data set includes several outliers. The last column "na" shows that the synthetic data set has no missing prices.

Table 3. Summary Statistics for the Price for Period 1Q2008

Minimum	q1	Median	Mean	q3	Maximum	Na
31,400	188,600.00	249,200	282,320	331,600	1,799,900	0

Table 4. Summary Statistics for the Floor Area for Period 1Q2008

Minimum	q1	Median	Mean	q3	maximum	Na
43	90.00	112	120	139	313	0

Note: q1 – First quartile q3 – Third quartile.

48.   The large difference between the minimum house price and maximum house price is an indication that outliers may be present in the data. Given that the median price is 249,200 and the interquartile range is 188,600-331,600, the price of the dwelling with the maximum value should be investigated for accuracy.

49.   **Verify the date field** – The compiler should ensure that the date field is properly formatted and that the date range is as expected. For example, if the compiler is only expecting prices for the first quarter of 2008 and finds observations dated the fourth quarter of 2008, further investigation may be required.

50.   **Check the number of occurrences for each variable** – Compilers may also want to produce a frequency count for the occurrence of each variable. The aim is to find the variables that can influence the price. For example, if only one dwelling in the data set has a garage then the compiler may not want to use this information when constructing the pricing model since it is not statistically representative. Frequency counts can also be useful when deciding how to group or stratify the data. For example, Table 5 shows the frequency of occurrences for the variable BER. Given the small number of observations in Classes A, E and F the compiler may need to consider aggregating the classes into a higher level as shown in Table 6 where A,B and C are aggregated together and D,E and F are aggregated together.

Table 5. Frequency of Occurrences for the Variable BER for Period 1Q2008

A	B	C	D	E	F
15	547	1,010	367	257	314

Table 6. Aggregations for the Variable BER for Period 1Q2008

A+B+C	D+E+F
1322	747

51.   **Analyze the coverage** – The compiler should also determine the coverage of the data sources (e.g. geographic coverage, type of dwellings, financing, etc.). For example, if the target coverage is only the urban areas, transactions in regions other than the urban should be removed. Table 7 is an example of a coverage report for the variable "Neighborhood_Type".

Table 7. Frequency of "Neighborhood Type for Period 1Q2008"

Neighborhood_Type	Frequency
Rural	466
Urban	2,055

52.   Check for duplicated records - Compilers should include a check for duplicate records in cases where duplicate records would be clear errors and may impact the subsequent processing of the data. For example, it may be logically incorrect for a dwelling with the same selling date and address to appear on the file multiple times. A validation report could be created which checks for cases where two observations have the same date and address. This type of verification is especially important when using web scraping techniques since the same dwelling can be listed by different real estate agents and by the owner on multiple websites. The synthetic data set that accompanies this *Guide* does not contain duplicated records.

53.   Check for missing values - Compilers may want to perform a check on missing values. One way to assess the extent of missing values is to calculate the percentage of missing values to total expected values. For example, the synthetic data set contains 332 missing values for the period 1Q2008. This represents 12.39 percent of the total number of observations. Table 8 is an example of a validation report that checks for missing values.

Table 8. Missing Values for Period 1Q2008

Period	1Q2008	1Q2008	1Q2008		1Q2008	1Q2008	1Q2008
id	30	46	55		3541	3560	3570
Year	2008	2008	2008		2008	2008	2008
Month	1	1	1		3	3	3
Dwelling_Type	Detached	Terraced	NA		NA	NA	Terraced
Price	789200	202100	354100		205000	172500	281900
Floor_Area	296	129	125		76	141	111
Neighborhood_Affluence	32.1	23.4	31.6	(…)	5.7	6.9	22.7
Year_Built	2000-2009	1910-1929	1950-1969		1930-1949	1930-1949	1930-1949
BER	C	C	F		E	D	C
County	62	71	72		71	84	71
Region	6	7	7		7	8	7
Neighborhood_Type	Urban	Urban	Urban		Urban	Urban	Urban
Building_Levels	NA	NA	NA		NA	NA	NA
Central_Heating	Mains_Gas	Mains_Gas	Mains_Gas		Mains_Gas	Mains_Gas	Mains_Gas
Status	Existing	Existing	Existing		Existing	Existing	Existing

Box 5. Imputations

When compiling an RPPI it is often necessary to impute values when they are missing or not reported. For price indices where the matched model is used, this is particularly important since a price is required for each observation in each time period.

If imputations are required three options are suggested: impute with the average if the compiler is imputing a numerical value (such as the price), with the mode if the variable is qualitative (e.g., the unit has a garage), or use expert knowledge to determine the most likely value. Regardless of the technique selected, a degree of bias is introduced in the index when estimated prices are used. The compiler should always seek to obtain as much observed data as possible.

54.   Identify errors and outliers - While it is straightforward to identify missing values, identifying errors and outliers is more challenging. First, it is important to distinguish errors from outliers. While an outlier is plausible but not very probable, an error is a value that is not plausible. For example, a 2,500 square meter single-family house with 50 bedrooms is plausible but not very probable, whereas an apartment with 2,000 bedrooms is not plausible and would be an error. Errors can occur because of data entry errors or mis-reported information. Compilers should establish error and outlier detection routines that are generated each time a new file is received from the data provider.

**55.   **One way to identify outliers is to determine a plausible range of values and track all observations that fall outside of the range. This range can either be developed using a variety of statistical techniques or based on the complier's judgement and knowledge of the industry. The limits used should be defined according to the data characteristics of each country and for each stratum. Different limits should be tested before setting the final values to be used in the outlier detection routine. Once limits are determined, they should be used for at least one year.

**56.   **In this exercise the summary statistics shown in Table 9 will be used to define the plausible range of values. The lower limit of the range is the first quartile minus the difference interquartile (q1- (q3-q1)) and the upper limit of the range is the sum of the third quartile with the difference interquartile (q3+(q3-q1)). These limits are amplified by multiplying them by a factor. The factor is a number somewhere between 0.5 and 2, that is not the same for every country. The compiler should test the results for different values assumed by the factor.

Table 9. Summary Statistics for the Price per Area for Period 1Q2008

Minimum	q1	Median	Mean	q3	Maximum
131	1,703.00	2,307	2,488	3,066	16,564

**57.   **Using the above technique 120 records were identified as outliers due to the extreme price per floor area (P_area) as show in Table 10.

Table 10. List of Outliers

Period	1Q2008	1Q2008	1Q2008		1Q2008	1Q2008	1Q2008
Id	17	18	37		3463	3476	3543
Year	2008	2008	2008		2008	2008	2008
Month	1	1	1		3	3	3
Dwelling_Type	Detached	Semi-Detached	Apartment		Apartment	Terraced	Semi-Detached
Price	634600	429800	227600		245500	657800	271900
Floor_Area	134	93	44		46	115	44
Neighborhood_Affluence	29.2	26.6	25.4	(..)	40.3	29.9	33.1
Year_Built	1970-1979	1980-1989	1980-1989		2000-2009	2000-2009	1980-1989
BER	F	D	F		C	C	E
County	72	72	71		71	71	71
Region	7	7	7		7	7	7
Neighborhood_Type	Urban	Urban	Urban		Urban	Urban	Urban
Building_Levels	2_or_more	2_or_more	2_or_more		2_or_more	2_or_more	1
Central_Heating	Oil	Mains_Gas	Electricity		Mains_Gas	Mains_Gas	Mains_Gas
Status	Existing	Existing	Existing		Existing	Existing	Existing
P_area	4735.82	4621.50	5172.73		5336.95	5720	6179.55

58. One practical way to visualize the existence of outliers is with the use of histograms. Figure 3 shows histograms for price per floor area ("P_area") and per vintage (New, Existing). Given the large tails it is clear there are several outliers in the dataset.

Figure 3. Histograms Before Cleaning the Data
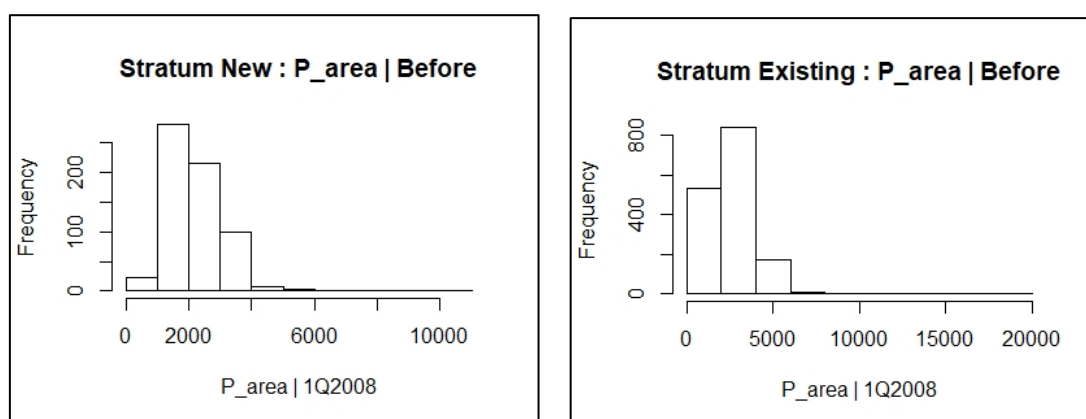
59. The results after data cleaning, i.e. deleting duplicates, missing and outliers, can be seen in Figure 4. The histograms no longer show a long right tale.

Figure 4. Histograms After Cleaning the Data
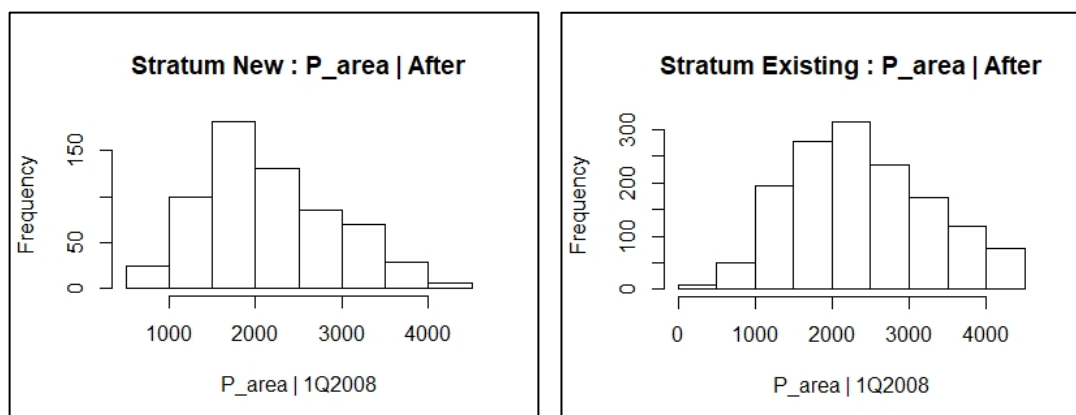
60. Other types of thresholds can be established to control for outliers. The following tables from O'Hanlon (2011) show examples of how establishing boundaries or ranges can be used to identify outliers. These boundaries should be adapted to the circumstances of the housing market of each country. The country data should be analyzed to understand which values should be used for these boundaries. For instance, a compiler could establish a range of bedrooms that could be associated with each property type. If an observation falls outside of the range, it would be flagged as an outlier. Table 11, Table 12, and Table 13 illustrate three types of thresholds that could be established when attempting to detect outliers within a set of residential property data.

Table 11. Example for Threshold of Number of Bedrooms by Property Type

Property Type	Allowable Range for Number of Bedrooms
Detached House	between 2 and 6
Semi-detached House	between 2 and 5
Bungalow	between 1 and 6
Terraced House	between 1 and 5
Apartment/Flat	between 1 and 4

Table 12. Example of Thresholds for Floor Area by Property Type and Bedrooms

Property Type/Number of Bedrooms	Allowable Range for Floor area (meters squared)					
	1	2	3	4	5	6
Detached House	-	40-239	60-279	100-339	120-419	120-479
Semi-detached House	-	40-179	60-199	80-219	80-339	-
Bungalow	10-219	20-219	40-239	80-319	100-379	100-419
Terraced House	20-159	40-159	60-159	60-259	80-379	-
Apartment/Flat	20-119	40-139	40-159	40-219	-	-

Table 13. Example of Thresholds for Number of Rooms by Property Type and Bedrooms

	Allowable Range for Number of Rooms					
Detached House	-	4-8	5-12	6-14	7-16	8-17
Semi-detached House	-	4-8	5-10	6-12	7-14	-
Bungalow	3-8	4-8	5-10	6-13	7-15	8-15
Terraced House	3-6	4-7	5-10	6-12	7-13	-
Apartment/Flat	3-5	4-7	5-8	6-10	-	-

C. Weights

Reference: Handbook on Residential Property Price Indices, pages 39 and 156

61.   Weights play an important role in the compilation of a price index. Weights are used to aggregate prices from the primary level into more meaningful aggregates and groups. In the case of residential properties, the primary levels are usually equated to a stratum. Strata are useful ways to group like items together.

62.   The choice and quality of weights can have a significant bearing on any price index calculation. For example, using a mean of prices to create an aggregated index assumes that each price is of equal importance. In most cases this is not realistic. There are items in the basket that are of higher importance than other items. As such, the weights should be differentiated and correspond to the importance of the item in the basket.

63.   There are two types of weights that an RPPI compiler can use: stock weights and flow weights. Stocks are positions in time and flows reflect changes from one period to the next. For example, a stock weight could be the number of properties in a geographic region on a specific date. A flow weight could be the number of properties that were sold during a certain period (one year is recommended). Flow weights are therefore associated with the transactions that took place in a certain period. Stock weights are associated with the accumulated flow of transactions that have taken place from a point in time. The flow weights are suited for monitoring financial stability and the stock weights are more appropriate if the index is being used to track the price change in the stock of housing.

64.   In most countries stock weights are derived from census data. This presents several challenges for the compiler. First, the census may not collect the level of detail required by the compiler such as information related to the number of bedrooms, bathrooms, and area. Second, a census is conducted on a periodic basis, often every 10 years and the weights may become outdated.

65.   Flow weights are generally derived from administrative records such as real estate transactions or loan data obtained from financial institutions. Generally, when an administrative

data source[5] is used all transactions are available and weights for each transaction type can be estimated. In some cases, a census of administrative transactions cannot be obtained, and a sample of the transactions is only available. When a sample is used (for example, data from web scraping) a complementary data source is required to estimate the weights otherwise it is assumed that the mix of dwellings in the sample is representative of the mix of dwellings being transacted in that period. The weights can be derived from one year of data or from an average of years (three is the general rule) to obtain more robust results. As a best practice the weights should be updated annually and kept constant for the year when constructing a sub-annual index.

66. Where values are not available for weights, volumes (e.g., number of transactions, square feet, etc.) are an adequate substitute.

67. The following table illustrates weights calculated from the synthetic data provided with this *Guide.* The files with data for 2008 and the R script "weights.r" are used in this example. Note that the calculation of weights should be done after the data have been "cleaned".

Table 14. Weights Compilation – Summing the Transactions Value for the New Dwellings Status

Period	1Q2008	1Q2008	1Q2008		4Q2008	4Q2008	4Q2008
id	6	22	27		17886	17892	17896
Year	2008	2008	2008		2008	2008	2008
Month	1	1	1		12	12	12
Dwelling_Type	Terraced	Detached	Terraced		Terraced	Detached	Semi_Detached
*Price**	*318400*	*216500*	*205500*		*204900*	*618500*	*168300*
Floor_Area	94	152	86		101	276	135
Neighborhood_Affluence	35	18	25.4	(	27.4	38.6	28.2
Year_Built	2000_2009	2000_2009	2010_or_more	…	2010_or_more	2000_2009	2000_2009
BER	High	Medium	High	)	High	High	High
County	62	63	74		84	12	81
Region	6	6	7		8	1	8
Neighborhood_Type	Urban	Rural	Urban		Urban	Urban	Urban
Building_Levels	2_or_more	1	1		2_or_more	1	2_or_more
Central_Heating	Mains_Gas	Oil	Mains_Gas		Mains_Gas	Other	Mains_Gas

- ***SUM of Prices for New Dwellings during 2008 = 626,111,900***

---

[5] Administrative data sources refer to lawyers, notaries, land registries, tax authorities, or municipalities.

Table 15. Weights Compilation – Summing the Transactions Value for the Existing Dwellings Status

Period	1Q2008	1Q2008	1Q2008		4Q2008	4Q2008	4Q2008
id	1	2	3		17893	17894	17895
Year	2008	2008	2008		2008	2008	2008
Month	1	1	1		12	12	12
Dwelling_Type	Semi_Detached	Apartment	Detached		Terraced	Semi_Detached	Detached
Price	305300	106500	386500		171700	274500	194200
Floor_Area	109	43	147		86	95	156
Neighborhood_Affluence	44.6	8.7	24.3		18.3	28.5	4.8
Year_Built	2000_2009	1990_1999	1980_1989	(...)	1970_1979	2000_2009	1990_1999
BER	Medium	Low	Medium		Medium	Medium	Medium
County	62	41	41		11	62	34
Region	6	4	4		1	6	3
Neighborhood_Type	Urban	Urban	Urban		Urban	Urban	Rural
Building_Levels	2_or_more	1	2_or_more		2_or_more	2_or_more	1
Central_Heating	Mains_Gas	Electricity	Oil		Oil	Mains_Gas	Oil

SUM of Prices for Existing Dwellings during 2008 = 2,005,758,600

68.   Second, the stratum weights are derived by dividing the total value of all properties in the stratum by the total value of all properties as shown in Table 16. The conclusion one can draw from the weights is that 24 percent of the dwellings transacted during 2008 are new dwellings and the remaining 76 percent are existing dwellings (i.e. the data are heavily weighted towards existing dwellings). In the above example, values were used to derive the weights. Compilers may want to experiment with using other (quantity based) information to derive the weights.

Table 16. Weights Compilation for Year 2008

Total_W	1
W_New	626111900 / 2631870500 = **0.24**
W_Existing	2005758600 / 2631870500 = **0.76**

METHODS

69.   This *Guide* illustrates the two most commonly used methods to compile an RPPI: (1) stratification with simple averages and (2) hedonic methods. In the case of hedonic methods there are three variants that are used by compilers: (a) Time dummy, (b) Imputations, and (c) Characteristics. A summary of the various methods presented in this *Guide* are presented in Table 1.

70.   As a general rule, the lower the number of observations, the more likely it is that the time dummy method with a rolling window [6]will be most appropriate, as the time dummy variables are used for the index, rather than relying on weighted averages of all the other coefficients. The larger the dataset, the more likely it is that the imputations or characteristics methods can be used.

Table 17. Methods for RPPI Compilation

Methods	Name	Used if...
1	Median with stratification	... prices and at least one characteristic, (e.g., location) are available.
2	Time dummy hedonic method	... few price observations per period are available or a monthly index is required.
3	Imputations hedonic method	... more detailed data are available. It can also be used with geospatial data.[7]
4	Characteristics hedonic method	... more detailed data are available. Equivalent to the Imputations hedonic method when the model is log-linear and geospatial data are not used.

71.   While this Guide illustrates the most commonly used methods to compile an RPPI there are two methods that the *Guide* will not address: the repeated sales method and the sales price appraisal ratio (SPAR) method.

72.   A well-known example of the repeated sales method is the Case-Shiller Index, which matches dwellings when they are re-sold. Despite the transparency in the calculations it is not covered in this *Guide* for the following reasons:

•   It is very data intensive, particularly in terms of historical data. Therefore, it is difficult to

---

[6] Rolling window is a method that replaces data from prior periods with data from more recent periods. For example, assume the rolling window is 12 months and assume the compiler has 12 months of data from January 2019 to December 2020. When the compiler processes January 2020 – data from January 2019 is removed from the model.

[7] Geospatial data or geographic information are the data or information that identify the geographic location of features and boundaries on Earth, such as natural or constructed features, oceans, and more. Spatial data are usually stored as coordinates and topology.

implement in many countries.

- If it does not consider depreciation and renovations, it will lead to a bias in the resulting index.

- It may suffer from sample bias since the mix of dwellings taken in the calculations for each period may not be representative.

- Dwellings sold for the first time are not covered.

- The published index value for any period is subject to constant revisions as new data involving a property transacted in that period and later periods become available.

73. The SPAR method matches the transaction value with the appraisal value from a previous period. This method is applicable in countries where the properties are reassessed frequently, normally for tax purposes.

D. Method 1: Median with Stratification

Reference: Handbook on Residential Property Price Indices, pages 38- 48

74. One approach that could be taken to measure the change in residential property prices is to compare the average price of properties transacted in a given period (period t) to the average prices of properties transacted in a base or reference period (period 0).[8] While this approach is straightforward and easy to implement it is discouraged due to the extreme heterogeneity of properties. This is a second-best approach and should only be used when the information available to the compiler is limited to the price and date of the transaction.

Box 6. Should Mean or Median be Used?

Given house price distribution is often positively skewed the median price is likely the best estimate of the price of the "typical" property. The use of the median incorporates the assumption that prices for all dwellings will follow the price trend of the typical dwelling. The median is used more widely in RPPI compilation than the mean as it is less subject to price fluctuations that can be caused by a small number of high-priced properties. Also, by use of the geometric mean, these concerns can often be allayed. More detail on this discussion can be found on the *Handbook on Residential Property Price Indices, pages 38-48.*

---

75. When compilers have access to only one variable such as the location, the size, or the type of property the stratification method can be used. Stratification is a technique by which the compiler groups observations together according to some common set of characteristics. For example, if a set of data contains information about single dwellings as well as apartments it

---

[8] Averages are measures of central tendency with the most measures common being median, mean, and mode.

would be recommended to create two strata, one that includes all single dwellings and one that includes all apartments.

76.   While stratification will help control for the quality-mix between strata the mix of dwellings within a stratum will still tend to be rather heterogeneous. The compiler may be tempted to increase the stratification levels for example by city and within each city by type of dwelling and further down by building energy rating, etc. This could easily result in a situation where the compilers would end up with more strata than data points. Compilers must balance the need for detailed strata with the need to have enough observations in each stratum to calculate an accurate index. If the stratum is too narrowly defined, there may not be enough observations to obtain a robust estimate in a given period.

77.   Building strata requires expert judgement and data analysis and depends in large part on the specificities of the available data set. While it is not possible to prescribe a precise and detailed method to build appropriate strata some general guidelines have emerged.[9]

78.   There is a critical distinction between strata, used in the stratification method, and dimensions of the data, which are used in the regression method (explained in the following chapters). For example, while a data set will generally not have enough observations to create a stratum for five-bedroom two-bathroom detached properties between 10 and 19 years old, a regression model could readily take advantage of these characteristics.

79.   Location is generally the first level of stratification. The price of a dwelling is mainly influenced by its location. In general, RPPI compilers have access to the price, date of the transaction, and address of the property. The address often conforms to administrative boundaries (e.g., county) and can be a good starting point to develop strata. For example, a stratum could be created for each town. If stratifying by counties does not provide enough observations the RPPI compiler could group counties together based on their similarities (e.g., population size). To find the similarities among the towns the compiler should consult with agents, developers, and other professionals working in the real estate market.

80.   RPPI compilers can also stratify their data by the vintage of the dwelling (new or existing). New and existing dwellings are different market segments with price trends that might differ. Also, the variables that influence the prices of new and existing dwellings are not the same. For new dwellings it may be important to use a "state of completion" variable in countries where dwellings are sold to households before or after completion. For existing dwellings, a variable identifying the renovation status may be important.

81.   User needs should also be taken into consideration when developing the stratification strategy. The compilation of sub-indices is a by-product of the stratification process. These sub-

---

[9] The number of strata should be informed by an examination of the data. For example, a compiler may want to select strata and then calculate the standard error of average prices to determine if the proposed strata are appropriate.

indices may have economic meaning and be useful for analysis and data validation.

82. Table 18 shows an example of an RPPI with strata based on the vintages of the dwellings. The first stratum includes observations of new dwellings and the second stratum includes observations of existing dwellings. Sub-indices are compiled at the level of strata.

Table 18. Stratification – Index Structure

Level	1Q2008	2Q2008	3Q2008	4Q2008
RPPI	105.0	103.3	98.3	93.4
New	106.9	101.7	96.2	95.2
Existing	104.4	103.8	98.9	92.9

To help compilers become familiar with the process of calculating an RPPI using the median with stratification method we encourage them to complete exercise 2 which has been developed to accompany this *Guide*.

Exercise 2: Calculation of a Quarterly RPPI Using a Median with Stratification

83. The index compilation for the first year prior to publication follows the next four steps. The calculations can be replicated using the R script named "Stratification_starting.r" and the synthetic data sets that accompanies this *Guide*. The R script includes guidance on how to adapt it to other data sets.

84. The strata will be defined as in Table 18, with Stratum 1 including all new dwellings and Stratum 2 including all existing dwellings.

85. **Step 1 – Calculate the median** - After strata have been defined the median price within each stratum is calculated for each period. In a first stage, the first quarter of 2008 is the reference period. In Table 19 we can see that the median price for new dwellings is 231,900 and for exiting dwellings is 242,300. This will be the base price used to calculate the indices in the first year.

Table 19. Stratification – Calculation of the Median in the Reference Period (1Q2008)

Dwelling_Type	Price	Floor_Area	Neig_Afflu	Year_Built	BER	County	Region
Semi_Detached	305300	109	44.6	2000_2009	Medium	62	6
Apartment	106500	43	8.7	1990_1999	Low	41	4
Detached	386500	147	24.3	1980_1989	Medium	41	4
(…)							
Semi_Detached	393800	96	18.8	2000_2009	Medium	73	7
Semi_Detached	168100	129	19.1	1990_1999	Medium	54	5
Detached	751500	208	21.7	1950_1969	Low	71	7
Median New							**213900**
Terraced	318400	94	35	2000_2009	High	62	6
Detached	216500	152	18	2000_2009	Medium	63	6
Terraced	205500	86	25.4	2010_or_more	High	74	7
(…)							
Terraced	158400	94	27.3	2000_2009	Medium	42	4
Terraced	135800	82	32.1	2000_2009	Medium	22	2
Terraced	273900	81	34.8	2000_2009	High	44	4
Median Existing							**242300**

Table 20. Stratification – Calculation of the Median in the Current Period (2Q2008)

Dwelling_Type	Price	Floor_Area	Neig_Afflu	Year_Built	BER	County	Region
Terraced	201300	93	23.2	2000_2009	High	85	8
Apartment	137500	72	17.1	2000_2009	Medium	41	4
Detached	506000	161	35.6	2000_2009	Medium	44	4
(…)							
Semi_Detached	215800	120	17.9	2000_2009	High	62	6
Detached	247400	139	15.7	2000_2009	Medium	43	4
Detached	461900	133	21.4	2000_2009	High	41	4
Median New							**220600**
Semi_Detached	385300	133	31.5	1990_1999	Medium	73	7
Semi_Detached	308900	119	27.8	2000_2009	Medium	74	7
Semi_Detached	472000	116	29.7	1990_1999	Medium	72	7
(…)							
Terraced	136500	83	7.8	1950_1969	Low	83	8
Detached	817100	262	19.3	2000_2009	Medium	73	7
Semi_Detached	261400	91	23.6	1990_1999	Medium	74	7
Median Existing							**241000**

86. **Step 2 – Calculate the sub-indices** - The sub-indices are compiled by dividing the median price in the current period (t) with the median price in the reference period (0) for each stratum. The example in Table 21 shows the calculation of the sub-indices for 2Q2008. Both prices decreased. New dwellings decreased 4.9 percent compared to the first quarter, while existing dwellings shown more stability with prices decreasing only by 0.5 percent since the first quarter.

Table 21. Stratification – Compilation of the Sub-Indices

	Median 2008Q1	Median 2008Q2	Calculation	Sub-index
New	231900	220600	220600 / 231900	95.1
Existing	242300	241000	241000 / 242300	99.5

* User may not obtain the same result due to rounding.

87. **Step 3 – Aggregation (Laspeyres)** – The sub-indices are then aggregated by a weighted average using the base year weights. In the context of this exercise the base year weights represent the total value of the properties classified in the stratum divided by the total value of all properties in all strata as calculated in the weights exercise. The indices for 1Q2008, the base period, are 100 by definition. The index for the RPPI for the second quarter of 2008 decreased by 1.6 percent compared to the first quarter.

Table 22. Stratification – Aggregation of the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Weights
RPPI	100.0	98.4	93.6	89.0	1.0
New	100.0	95.1	90.0	89.0	0.24
Existing	100.0	99.5	94.7	88.9	0.76

88. **Step 4 – Re-reference the sub-indices** – At this point the reference period of the sub-indices corresponds to the first period (1Q2008). However, the reference period should be a full year and not a quarter. In order to re-reference the sub-indices to the full year (2008=100) the index value for each quarter is divided by the annual average index value (calculated as the mean of the quarterly indices). The resulting average of the four re-referenced quarters will be 100. The RPPI in the 1Q2008 was 5.0 percent higher than the annual average for 2008.

Table 23. Stratification – Re-Referencing the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Re-Referencing Factor (Mean of Indices)
RPPI	100 / 0.95 = 105.0	98.4 / 0.95 = 103.3	93.6 / 0.95 = 98.3	89.0 / 0.95 = 98.3	(100 + 98.4 + 93.6 + 89) / 4 / 100 = 0.95
New	100 / 0.94 = 106.9	95.1 / 0.94 = 101.7	90.0 / 0.94 = 96.2	89.0 / 0.94 = 95.2	(100 + 95.1 + 90.0 + 89.0) / 4 / 100 = 0.94
Existing	100 / 0.96 = 104.4	99.5 / 0.96 = 103.8	94.7 / 0.96 = 98.9	88.9 / 0.96 = 92.9	(100 + 99.5 + 94.7 + 88.9) / 4 / 100 = 0.96

* User may not obtain the same result due to rounding.

89.   The index compilation for the following years is as described in the next five steps. The calculations can be replicated using the R script named "Stratification_production.r." In addition, a base price should be calculated every year. The R script named "Base price.r" can be used to perform this calculation. The scripts include guidance on how they can be adapted to other data sets.

90.   **Step 1 – Calculate the reference price** - The reference price is the median of the transaction prices in the reference period. In this *Guide* the indices will be compiled in steps. The first step is to calculate the median price of the last quarter of the previous year (the median prices for 4Q2008 are shown in Table 24).

Table 24. Stratification – Reference Prices

	4Q2008
New	206500
Existing	215500

91.   **Step 2 – Calculate the median for the current period** – Second, the median price for the current period (1Q2009) is calculated as shown in Table 25. The calculation was done use the R script: "Stratification_production.r."

Table 25. Stratification – Median Prices in the Current Period (1Q2009)

	1Q2009
New	195000
Existing	203350

92.   **Step 3 – Compilation of the sub-indices** - The sub-indices are then compiled by dividing the median of the current period price (1Q2009) by the median of the reference period price (4Q2008) giving the price change in 1Q2009 compared to 4Q2008. For example, prices forboth new and existing dwellings decreased 5.6 percent since 4Q2008.

Table 26. Stratification – Compilation of the Sub-Indices (1Q2009)

Sub-Indices	1Q2009
New	195000 / 206500 * 100 = 94.4
Existing	203350 / 215500 * 100 = 94.4

* User may not obtain the same result due to rounding.

93. **Step 4 – Aggregation (Laspeyres)** – The sub-indices are then aggregated using the base year weights.

Table 27. Stratification – Aggregation of the Sub-Indices

Level	1Q2009	Weights
RPPI	94.4	1.0
New	94.4	0.24
Existing	94.4	0.76

Box 7. Chained Indices

In the example that accompanies this *Guide,* indices are first compiled based on the last quarter of the previous year. However, the reference period of the indices to be published is a full year (2008). To bring the index reference period to 2008, indices are chained. This approach allows for annual updates of the index weights and the base price or base average characteristics as explained later in this *Guide*.

The formula below shows how to obtain the index of 3Q2010 in base 2008 from the 3Q2010 index in base 4Q2009.

$$I_{2008}^{3Q2010} = I_{4Q2009}^{3Q2010} * I_{2008}^{4Q2009}$$

Let's suppose, $I_{4Q2009}^{3Q2010} = 95.9$, $I_{2008}^{4Q2009} = 101.8$ and $I_{2008}^{3Q2010} = 107.3$ the result would be 95.9 * 107.3 /100 = 102.8.

While 95.9 shows a price decrease of 4.1 percent since the fourth quarter of 2009, the index 102.8 shows a price increase of 2.8 percent since 2008.

Further guidance can be found in the Chapter 9 of the *CPI Manual: Concepts and Methods*.

---

94. **Step 5 – Chaining the indices** - All indices (stratum indices and the total RPPI) of the current period are chained (multiplied) by the chained index of the last period of the previous year. Table 26 shows an example of chaining the RPPI. The chained index is the final result. In this example, the chained index in 1Q2009 decreased 11.8 percent compared to the average of 2008.

Table 28. Stratification – Chaining the Indices

Quarter	Unchained Index	Calculation	Chained Index
1Q2008			105.0
2Q2008			103.3
3Q2008			98.3
4Q2008			**93.4**
1Q2009	94.8	94.8 * 93.4 / 100	88.2
2Q2009	92.4	92.4 * 93.4 / 100	86.3
3Q2009	87.1	87.1 * 93.4 / 100	81.3
4Q2009	82.3	82.3 * 93.4 / 100	**76.9**
1Q2010	91.9	91.9 * 76.9 / 100	70.7
2Q2010	93.0	93.0 * 76.9 / 100	71.5
3Q2010	93.3	93.3 * 76.9 / 100	71.7
4Q2010	91.0	91.0 * 76.9 / 100	**70.0**

* User may not obtain the same result due to rounding.

95. As explained before, the stratification process controls for quality differences of dwellings between strata, but not within. The need to control for quality-mix change within strata can be achieved using hedonic regressions within a stratum. In addition, it is possible to generate reliable sub-indices using hedonics that cannot be generated using stratified medians. For these reasons hedonic methods are preferred by most compilers.

E. Hedonic Methods

Reference: Handbook on Residential Property Price Indices, pages 50-51

96. To measure pure price changes, it is essential to measure the same product over time, adjusting for changes in quality and quantity. In the case of RPPIs, quality adjustments require detailed information about the characteristics of the properties and can be made using the hedonic method.

Box 8. Hedonic Regression

"The hedonic regression method recognizes that heterogeneous goods can be described by their attributes or characteristics. That is, a good is essentially a bundle of (performance) characteristics. In the housing context, this bundle may contain attributes of both the structure and the location of the properties. There is no market for characteristics, since they cannot be sold separately, so the prices of the characteristics are not independently observed. The demand and supply for the properties implicitly determine the characteristics' marginal contributions to the prices of the properties. Regression techniques can be used to estimate those marginal contributions or shadow prices." (*Handbook*)

---

97. Hedonic regressions recognize that a dwelling is composed of a bundle of characteristics. These characteristics are not sold separately, and consequently their prices cannot be independently observed in the selling price. Even though these prices are not observed they can be estimated with hedonic regression techniques. These are often referred to as "shadow" prices.

98. Several functional forms for hedonic regressions are possible and discussed in the *Handbook*. This *Guide* will use the log-linear form. The distribution of real estate prices is usually positively skewed. The log-linear form reduces the impact of skewness in the price distribution and consequently, reduces heteroscedasticity. The log-linear form allows for parabolic relationships of the variables and for a multiplicative association between characteristics.

99. The specification of a hedonic equation in the log-linear form is as follows:

$$ln\ p_n^t = \sum_{k=1}^{K} \beta_k Z_{nk}^t + \varepsilon_n^t \tag{1}$$

t - period
n – number of dwellings in period t
k – each of K characteristics
$ln\ p_n^t$ – price logarithm
β_k^t – "shadow" price of characteristic k in period t
Z_{nk}^t –the value of characteristic k in dwelling n in period t
ε_n^t – error term

100. For example, consider two houses located next to each other (A and B) transacted in two consecutive periods. Both houses are the same size (800 sqm), but house A has a balcony and house B does not. House A has 3 bathrooms while house B has 2 bathrooms. Both houses have two car garages. Assume that in period t house A sold for 465,000 and in period t+1 house B sold for 400,000. A very simple (and misrepresented) analysis of this data would be to conclude that housing prices in this neighborhood have fallen 16 percent. Comparing the prices of these houses is not recommended since the characteristics of house A are quite different from house B. Using bundles of observations with variation in these characteristics, hedonic regression allows the compiler to decompose the price of each house into component "shadow prices". This is shown in the following table.

Table 29. "Shadow" Prices of Houses A and B

Characteristics	House A Shadow Prices	House B Shadow Prices
Size	350,000	350,000
Bathrooms	45,000	30,000
Garage	20,000	20,000
Balcony	10,000	
Total	425,000	400,000

101. For hedonic regressions to reliably estimate coefficients for each of the variables included, there needs to be sufficient depth in the dataset at the level of each of the variables. This is known as 'degrees of freedom' in statistics. For example, if you have only 20 observations but include 25 different variables that could all affect property prices, it is both logically and statistically impossible to work out the effect of each of the 25 variables. If you have 20

observations and only 10 variables, this is effectively asking the dataset to generate reliable results with just two observations per variable. While there is no set minimum threshold for the number of observations per variable, established practice (as well as statistical theory) suggest that a minimum of around 20 observations per variable are required to enable the regression to estimate the effect of a variable. In practice, this level can vary from period to period, but it holds in most cases.

102.   The R-script "fit model" can be used to find the best model to use for the regression. The model selected should be kept for at least one year. Any change in the model is a fundamental methodological change that needs to be documented and communicated to users. To specify the model the R-script was run over several periods.

103.   The model is run with all the variables available in the data set using ANOVA (Analysis of variance) that takes the AIC (Akaike criterion) to determine the goodness of fit. The lower the AIC value the better the fit. The results for data pertaining to 1Q2008 are below:

Final Model:

log(Price) ~ Dwelling_TypeApartment + Dwelling_TypeDetached + Dwelling_TypeSemi_Detached + Floor_Area + Neighborhood_Affluence + BERHigh + Neighborhood_TypeRural + Central_HeatingMains_Gas + Year_Built_Agg1950_1979 + County_Agg11_14 + County_Agg22_31 + County_Agg42_52 + County_Agg53_63 + County_Agg71_81

Figure 5. Best fitted model for 1Q2018

```
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                11.2774315  0.0510426 220.941  < 2e-16 ***
Dwelling_TypeApartment     -0.0720437  0.0268303  -2.685 0.007308 **
Dwelling_TypeDetached       0.0592598  0.0257980   2.297 0.021715 *
Dwelling_TypeSemi_Detached  0.0811982  0.0181106   4.483 7.75e-06 ***
Dwelling_TypeTerraced              NA         NA      NA       NA
Floor_Area                  0.0050861  0.0002085  24.394  < 2e-16 ***
Neighborhood_Affluence      0.0156396  0.0008853  17.666  < 2e-16 ***
BERHigh                    -0.0719893  0.0219078  -3.286 0.001033 **
BERLow                             NA         NA      NA       NA
Neighborhood_TypeRural     -0.0576104  0.0204707  -2.814 0.004935 **
Neighborhood_TypeUrban             NA         NA      NA       NA
Building_Levels1            0.0032262  0.0246421   0.131 0.895850
Building_Levels2_or_more           NA         NA      NA       NA
Central_HeatingElectricity -0.0085095  0.0452618  -0.188 0.850890
Central_HeatingMains_Gas    0.1393261  0.0384666   3.622 0.000299 ***
Central_HeatingOil          0.0162189  0.0379149   0.428 0.668864
Central_HeatingOther               NA         NA      NA       NA
Year_Built_Agg1950_1979     0.0625627  0.0263597   2.373 0.017716 *
Year_Built_Agg1980_1999     0.0152364  0.0264950   0.575 0.565308
Year_Built_Agg2000_or_more -0.0116848  0.0292836  -0.399 0.689918
Year_Built_Aggor_less_1949         NA         NA      NA       NA
County_Agg11_14            -0.0750776  0.0347127  -2.163 0.030669 *
County_Agg22_31            -0.1219974  0.0364371  -3.348 0.000828 ***
County_Agg31_41            -0.0114438  0.0320609  -0.357 0.721175
County_Agg42_52             0.0418541  0.0320524   1.306 0.191767
County_Agg53_63             0.0362856  0.0308751   1.175 0.240035
County_Agg71_81             0.2790694  0.0309684   9.011  < 2e-16 ***
County_Agg82_85                    NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2935 on 2048 degrees of freedom
Multiple R-squared:  0.5488,    Adjusted R-squared:  0.5444
F-statistic: 124.5 on 20 and 2048 DF,  p-value: < 2.2e-16
```

104. The p-value estimates whether or not the results are due to random variation. The closer the p value is to zero the better, i.e., the less likely they are to be caused by random variations. The p-value of the regression in Figure 5 is very close to zero meaning that there is a very small likelihood that the result is due to random variation.

105. The R-squared reveals that 54.88 percent of the variability of the price logarithm is explained by our model. Normally this would be considered a low value however for house price indices this value is rather good. As noted earlier there are many housing characteristics that influence the price of a dwelling and it is unlikely that a country will have access to the broad range of variables required to achieve a high R-squared.

106. The F-statistic is used to evaluate the fitness of the model. Normally the higher the F-statistic, the better fitted our model is. however, this value should be analyzed with the table of the F-Statistics.

F. Method 2: Time Dummy Hedonic Method

Reference: Handbook on Residential Property Price Indices, pages 51-52

107.   The time dummy method is generally used when the compiler had access to a data set with a large number of characteristics but few transactions for each period. With this method, prices and characteristics (z_{nk}^t) of all dwellings (n) for several periods (t) are pooled in the same regression and a dummy variable is established for each period (D_n^t). The index follows directly from the estimated time dummy parameters.

108.   The log-linear specification is used, thus the model for regression is:

$$\ln p_n^t = \beta_0 + \sum_{t=1}^{T} \delta^t D_n^t + \sum_{k=1}^{K} \beta_k Z_{nk}^t + \varepsilon_n^t \tag{2}$$

t - period
n – number of dwellings in period t
k – characteristics
$\ln p_n^t$ – price logarithm
β_0 – intercept
δ^t – coefficient of the time dummy variable that will generate the index
D_n^t – time dummy variables
β_k^t – "shadow" price of characteristic k in period t
Z_{nk}^t – quantity of characteristic k in period t and dwelling n
ε_n^t – error term

109.   The index for current period (t) is derived as follows:

$$I_t = \exp(\hat{\delta}_t) * 100 \tag{3}$$

110.   For the reference period (0) a dummy variable is not required (the price index in the reference period is 100).

111.   If data from a new period are added, the indices from the previous period may change because the estimated coefficients $\hat{\delta}_t$ of the previous periods are revised based on the new observations. To avoid revisions to the indices a rolling window approach is used. Normally for quarterly indices the "shadow" prices of the characteristics ($\hat{\beta}$) are kept fixed for at least a year. This means that data from 12 months are used together. Every quarter a regression is estimated with data from the current quarter and the previous three quarters. The indices from the new 12-month window and the previous 12-month window are chained by using the last overlap period between the two windows.

112.   An example is given in the table below for a quarterly index:

Table 30. Example of Chaining the Indexes from Rolling Window Hedonic Regression

Quarter	Data from 1Q2010 to 4Q2010	Data from 2Q2010 to 1Q2019	Data from 3Q2010 to 2Q2019	Chained Index
1Q2010	**100.0**			**100.0**
2Q2010	101.2	**100.0**		101.2
3Q2010	101.1	101.3	**100.0**	101.1
4Q2010	100.9	101.2	100.8	100.9
1Q2019		101.0	100.3	100.7 = 100.9/101.2x101.0
2Q2019			101.4	101.8 = 100.7/100.3x101.4

* User may not obtain the same result due to rounding.

113. Ireland was one of the first countries to implement the time dummy hedonic method with a rolling window for the compilation of a house price index. O'Hanlon (2011) recommends using data from a 12-month window for a monthly index compilation. The 12-month period mitigates against all seasonality effects. A larger window of 15 months may also be used.

114. To help compilers become familiar with the process of calculating an RPPI using the time dummy hedonic method an exercise has been developed to accompany this *Guide*. The exercise (Exercise 3) uses a synthetic data set and a set of procedures written in R. The following step-by-step instructions are intended to walk the reader through the exercise.

Exercise 3: Step-by-step Calculation for a Time Dummy Hedonic Quarterly Index

115. The index compilation for the first year is presented in the seven steps noted below. The calculations can be replicated using the R procedure "time dummy_starting.r" and the synthetic data sets provided with this *Guide*.

116. Step 1 - Obtain data pertaining to four quarters.

117. Step 2 - Generate dummy variables - Create dummy variables for each period and for the categorical variables. Table 31 shows all the variables that will be used in the regression. The categorical variables are transformed in dummies namely the variable period. The first observation is an urban semi-detached dwelling sold in 1Q2008 with 109 square meters, built between 2000 and 2009, located in a building with 2 or more levels, in county 62 and in the region 6, heated with gas.

Table 31. Time Dummy – Data Set with the Dummy Variables

Variable	Obs 1	Obs 2	Obs 3	Obs (...)	Obs 4	Obs 5	Obs 6
Period1Q2008	1	1	1	(...)	0	0	0
Period2Q2008	0	0	0	(...)	1	1	1
Period3Q2008	0	0	0	(...)	0	0	0
Period4Q2008	0	0	0	(...)	0	0	0
Dwelling_TypeApartment	0	1	0	(...)	0	0	0
Dwelling_TypeDetached	0	0	1	(...)	0	0	0
Dwelling_TypeSemi_Detached	1	0	0	(...)	1	1	1
Dwelling_TypeTerraced	0	0	0	(...)	0	0	0
Floor_Area	109	43	147	(...)	133	119	116
Neighborhood_Affluence	44.6	8.7	24.3	(...)	31.5	27.8	29.7
Year_Built1910_1929	0	0	0	(...)	0	0	0
Year_Built1930_1949	0	0	0	(...)	0	0	0
Year_Built1950_1969	0	0	0	(...)	0	0	0
Year_Built1970_1979	0	0	0	(...)	0	0	0
Year_Built1980_1989	0	0	1	(...)	0	0	0
Year_Built1990_1999	0	1	0	(...)	1	0	1
Year_Built2000_2009	1	0	0	(...)	0	1	0
Year_Built2010_or_more	0	0	0	(...)	0	0	0
Year_Builtor_less_1910	0	0	0	(...)	0	0	0
BERHigh	0	0	0	(...)	0	0	0
BERLow	0	1	0	(...)	0	0	0
County11	0	0	0	(...)	0	0	0
County12	0	0	0	(...)	0	0	0
County13	0	0	0	(...)	0	0	0
County14	0	0	0	(...)	0	0	0
County21	0	0	0	(...)	0	0	0
County22	0	0	0	(...)	0	0	0
County23	0	0	0	(...)	0	0	0
County24	0	0	0	(...)	0	0	0
County25	0	0	0	(...)	0	0	0
County31	0	0	0	(...)	0	0	0
County32	0	0	0	(...)	0	0	0
County33	0	0	0	(...)	0	0	0
County34	0	0	0	(...)	0	0	0
County41	0	1	1	(...)	0	0	0
County42	0	0	0	(...)	0	0	0
County43	0	0	0	(...)	0	0	0
County44	0	0	0	(...)	0	0	0
County51	0	0	0	(...)	0	0	0
County52	0	0	0	(...)	0	0	0

Variable	Obs 1	Obs 2	Obs 3	Obs (...)	Obs 4	Obs 5	Obs 6
County53	0	0	0	(...)	0	0	0
County54	0	0	0	(...)	0	0	0
County61	0	0	0	(...)	0	0	0
County62	1	0	0	(...)	0	0	0
County63	0	0	0	(...)	0	0	0
County71	0	0	0	(...)	0	0	0
County72	0	0	0	(...)	0	0	1
County73	0	0	0	(...)	1	0	0
County74	0	0	0	(...)	0	1	0
County81	0	0	0	(...)	0	0	0
County82	0	0	0	(...)	0	0	0
County83	0	0	0	(...)	0	0	0
County84	0	0	0	(...)	0	0	0
County85	0	0	0	(...)	0	0	0
Region1	0	0	0	(...)	0	0	0
Region2	0	0	0	(...)	0	0	0
Region3	0	0	0	(...)	0	0	0
Region4	0	1	1	(...)	0	0	0
Region5	0	0	0	(...)	0	0	0
Region6	1	0	0	(...)	0	0	0
Region7	0	0	0	(...)	1	1	1
Region8	0	0	0	(...)	0	0	0
Neighborhood_TypeRural	0	0	0	(...)	0	0	0
Neighborhood_TypeUrban	1	1	1	(...)	1	1	1
Building_Levels1	0	1	0	(...)	0	0	0
Building_Levels2_or_more	1	0	1	(...)	1	1	1
Central_HeatingElectricity	0	1	0	(...)	0	0	0
Central_HeatingMains_Gas	1	0	0	(...)	1	1	1
Central_HeatingOil	0	0	1	(...)	0	0	0
Central_HeatingOther	0	0	0	(...)	0	0	0

118.   Step 3 - Build the strata (as per previous exercises).

119.   Step 4 – Calculate the "shadow" prices - Run a regression with the price logarithm as the dependent variable and all other variables as explanatory variables.

**120.   **There will be no coefficient for the first quarter. The index for the first quarter (reference period), equals 100 by definition.

Table 32. Time Dummy – OLS Results for New Dwellings

.rownames	Estimate	Std.Error	t.value
(Intercept)	11.82	0.06	194.34
Period2Q2008	-0.06	0.02	-3.74
Period3Q2008	-0.10	0.02	-6.09
Period4Q2008	-0.13	0.02	-7.67
Dwelling_TypeApartment	-0.03	0.02	-1.76
Dwelling_TypeDetached	0.10	0.02	4.42
Dwelling_TypeSemi_Detached	0.04	0.02	2.69
Floor_Area	0.00	0.00	22.49
Neighborhood_Affluence	0.01	0.00	14.07
BERHigh	-0.27	0.04	-6.11
Neighborhood_TypeRural	-0.05	0.02	-3.29
Building_Levels1	0.05	0.03	1.72
Central_HeatingElectricity	-0.04	0.03	-1.29
Central_HeatingMains_Gas	0.08	0.03	2.93
Central_HeatingOil	-0.08	0.03	-3.09
County_Agg11_14	-0.19	0.03	-6.86
County_Agg22_31	-0.15	0.03	-5.33
County_Agg31_41	0.00	0.03	0.05
County_Agg42_52	0.02	0.03	0.57
County_Agg53_63	-0.01	0.03	-0.33
County_Agg71_81	0.06	0.03	2.08

Table 33. Time Dummy – OLS Results for Existing Dwellings

.rownames	Estimate	Std.Error	t.value
Intercept)	11.10	0.03	428.57
Period2Q2008	-0.04	0.01	-3.41
Period3Q2008	-0.05	0.01	-4.93
Period4Q2008	-0.10	0.01	-9.87
Dwelling_TypeApartment	-0.13	0.02	-7.72
Dwelling_TypeDetached	0.15	0.01	10.60
Dwelling_TypeSemi_Detached	0.09	0.01	8.77
Floor_Area	0.00	0.00	39.70
Neighborhood_Affluence	0.02	0.00	44.08
BERHigh	-0.01	0.01	-1.21
Neighborhood_TypeRural	-0.15	0.01	-13.39
Building_Levels1	-0.03	0.01	-2.16
Central_HeatingElectricity	0.06	0.02	2.70
Central_HeatingMains_Gas	0.19	0.02	10.01
Central_HeatingOil	0.10	0.02	5.28
Year_Built_Agg1950_1979	0.08	0.01	6.68
Year_Built_Agg1980_1999	0.01	0.01	0.42
Year_Built_Agg2000_or_more	-0.09	0.01	-6.31
County_Agg11_14	0.04	0.02	2.02
County_Agg22_31	-0.09	0.02	-4.93
County_Agg31_41	0.05	0.02	2.85
County_Agg42_52	0.12	0.02	7.40
County_Agg53_63	0.08	0.02	5.04
County_Agg71_81	0.39	0.02	25.65

121. **Step 5 - Calculate the sub-indices** – Following from equation (3), δ are the coefficients of the time dummy variables. The coefficients of the other variables are now disregarded. Table 34 shows the results of these calculations, which show, for example, that that prices of new dwellings decreased 9.1 percent from 1Q2008 to 3Q2008.

Table 34. Time Dummy – Compilation of the Sub-Indices

Sub-Indices	1Q2008	2Q2008	3Q2008	4Q2008
New	100.0	exp (-0.06) * 100 = 94.2	exp (-0.10) * 100 = 90.8	exp (-0.13) * 100 = 88.2
Existing	100.0	exp (-0.04) * 100 = 96.5	exp (-0.05) * 100 = 95.2	exp (-0.10) * 100 = 90.3

* User may not obtain the same result due to rounding.

122. **Step 6 – Aggregation (Laspeyres)** - The aggregation of the sub-indices is a weighted average using property value weights (as explained in the first exercise).

Table 35. Time Dummy – Aggregation of the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Weights
RPPI	100.0	96.0	94.1	89.8	1
New	100.0	94.2	90.8	88.2	0.24
Existing	100.0	96.5	95.2	90.3	0.76

123.   Step 7 – Re-reference the sub-indices - The above indices are referenced to the first period (1Q2008). However, the reference period should be 2008=100. To re-reference the sub-indices such that 2008=100 the quarterly indices are divided by their mean. Table 36 shows, for example, that the RPPI in 4Q2008 decreased 10.2 percent compared to the average of 2008.

Table 36. Time Dummy – Re-Referencing the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Re-referencing Factor (Mean of Indices)
RPPI	100 / 0.950 = 105.3	96.0 / 0.950 = 101.0	94.1 / 0.950 = 99.1	89.8 / 0.950 = 94.6	(100 + 96.0 + 94.2 + 89.8) / 4 / 100 = 0.950
New	100 / 0.933 = 107.2	94.2 / 0.933 = 100.9	90.8 / 0.933 = 97.3	88.2 / 0.933 = 94.6	(100 + 94.2 + 90.8 + 88.2) / 4 / 100 = 0.933
Existing	100 / 0.955 = 104.7	96.5 / 0.955= 101.1	95.2 / 0.955 = 99.7	90.3 / 0.955 = 94.5	(100 + 96.4 + 94.6 + 89.8) / 4 / 100 = 0.955

* User may not obtain the same result due to rounding.

124.   The index compilation for the subsequent years is the same as described in the previous exercise except for Step 7, where the chaining occurs for the overall index and for all sub-indices. The calculations for the subsequent years (2009 and 2010) can be replicated by using the R procedure named "time dummy_production.r" and the synthetic data sets supplied with this *Guide*. Table 37 illustrates how to chain the overall index.

Table 37. Time Dummy – Chaining the Indices

Period	Indices from 1Q2008 to 4Q2008	Indices from 2Q2008 to 1Q2009	Indices from 3Q2008 to 2Q2009	Chained Index
1Q2008	**105.3**			**105.3**
2Q2008	101.0	**100.0**		101.0
3Q2008	99.1	98.2	**100.0**	99.1
4Q2008	94.6	93.7	95.5	94.6
1Q2009		88.1	89.6	88.9 = 94.6 / 93.7 x 88.1
2Q2009			86.4	85.7 = 88.9 / 89.6 x 86.4

* User may not obtain the same result due to rounding.

G.   Method 3: Characteristics Hedonic Method

Reference: Handbook on Residential Property Price Indices, pages 52-54

125.   The characteristics hedonic method measures the price evolution of a "typical" property. It relies on the use of a detailed set of property characteristics and is generally preferred to the

time dummy when the compiler has access to a large number of transactions for each period. The "typical" property is estimated by averaging the characteristics of all the properties in the stratum. For example, if there are 10 properties in a stratum and the average number of bedrooms among those 10 properties is four then the "typical" property would have four bedrooms.

126. A "shadow" price is then estimated for each characteristic in the current period (t) and in the reference period (0). The price index is calculated by comparing the price of the typical property in period (t) with the price of the "typical" property in the reference period (0).

Figure 6. The Mechanics of the Characteristics Hedonic Method

A property represents a bundle of characteristics → The "typical" property is a bundle of the average characteristics → Comparing the "prices" of period t with the "prices" of period 0, keeping the average characteristics constant → price variation of the "typical" property from period 0 to period t.
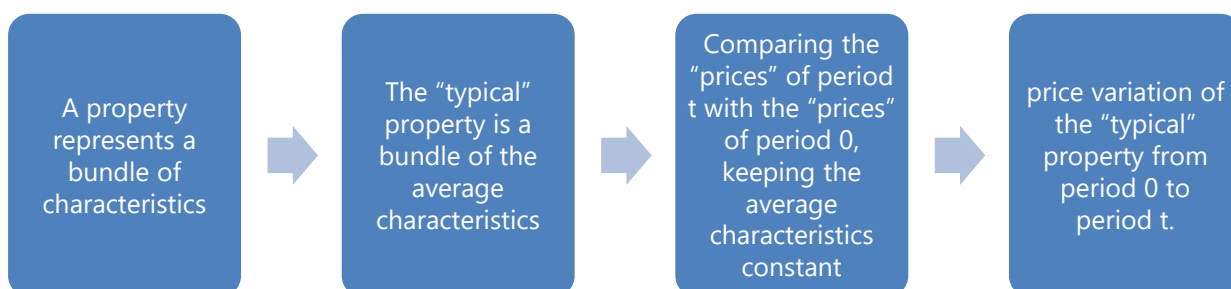
127. The mean is calculated for each numerical variable and the frequency for the categorical variables as in the example shown in the table below:

Table 38. Example of Calculation of the "Typical" Dwelling

	Dwelling A	Dwelling B	Dwelling C	Dwelling D	Dwelling E	Dwelling F	Mean/Frq (Typical Property)
Dwelling_TypeApartment	0	1	0	0	0	0	0.17
Dwelling_TypeDetached	0	0	1	0	0	0	0.17
Dwelling_TypeSemi_Detached	1	0	0	1	1	1	0.67
Dwelling_TypeTerraced	0	0	0	0	0	0	0.00
Central_HeatingElectricity	0	1	0	0	0	0	0.17
Floor_Area	109	43	147	133	119	116	111.17
Neighborhood_Affluence	44.6	8.7	24.3	31.5	27.8	29.7	27.77
Year_Builtor_less_1910	0	0	0	0	0	0	0.00
BERHigh	0	0	0	0	0	0	0.00
BERLow	0	1	0	0	0	0	0.17
County11	0	0	0	0	0	0	0.00
County12	0	0	0	0	0	0	0.00
County13	0	0	0	0	0	0	0.00
Neighborhood_TypeRural	0	0	0	0	0	0	0.00
Neighborhood_TypeUrban	1	1	1	0	0	0	0.50
Building_Levels1	0	1	0	1	1	1	0.67
Building_Levels2_or_more	1	0	1	0	0	0	0.33

128.   The log-linear specification for each stratum is used.

$$ln\ p_n^t = \beta_0 + \sum_{k=1}^{k} \beta_k z_{nk}^t + \varepsilon_n^t \tag{4}$$

129.   Separate regressions are run for this model, for each quarter in a stratum, using the data of the reference period (0) and the current period (t) to obtain the "shadow" prices. The "shadow" prices are the estimated regression coefficients (β').

130.   The index is obtained by exponentiating the difference between the estimated regression coefficients of the current period (t) and the reference period (0) multiplied by the average characteristics of the reference period (0).

$$I_0^t = \exp\left(\sum_{k=1}^{k}\left(\hat{\beta}_k^t - \hat{\beta}_k^0\right)\bar{Z}_k^0\right) \tag{5}$$

131.   To help compilers become familiar with the process to calculate an RPPI using the characteristics hedonic method an exercise has been developed to accompany this *Guide*. The exercise (Exercise 4) uses the synthetic data set and R procedures that accompany this *Guide*. The following step-by-step instructions are intended to walk the reader through the exercise.

Exercise 4: Step-by-step Calculation for a Quarterly Index with the Characteristics Hedonic Method

132.   The index compilation for the first year is presented in the following seven steps. The calculations can be replicated using the R procedure named "characteristics_starting.r" and the synthetic set of data that accompanies this *Guide*.

133.   **Step 1 - Obtain data**.

134.   **Step 2 - Build the strata** – As per exercise 1. The following steps are performed for each stratum.

135.   **Step 3 - Calculate the "shadow" prices** - Run a regression for each period with the price logarithm as the dependent variable and all other significant variables as the explanatory variables. Prior to running the regression transform the categorical variables into dummy variables. The estimates from a regression for 1Q2008 are shown in Table 39. The results can be interpreted as the "market-up" a given characteristic has on the overall price. For example, the fact that a dwelling is a detached house adds a 15 percent premium or mark-up to the house price.

Table 39. Characteristics – Coefficients for New Dwellings in the Reference Period (1Q2008)

.rownames	Estimate	Std.Error	t.value
(Intercept)	11.88	0.12	99.53
Dwelling_TypeApartment	0.00	0.03	0.12
Dwelling_TypeDetached	**0.15**	0.04	3.61
Dwelling_TypeSemi_Detached	0.14	0.03	4.73
Floor_Area	0.00	0.00	12.53
Neighborhood_Affluence	0.01	0.00	7.47
BERHigh	-0.28	0.09	-3.30
Neighborhood_TypeRural	-0.01	0.03	-0.48
Building_Levels1	0.08	0.05	1.67
Central_HeatingElectricity	-0.14	0.07	-1.98
Central_HeatingMains_Gas	0.07	0.06	1.12
Central_HeatingOil	-0.15	0.06	-2.49
County_Agg11_14	-0.32	0.05	-6.18
County_Agg22_31	-0.18	0.06	-3.34
County_Agg31_41	-0.12	0.05	-2.31
County_Agg42_52	-0.09	0.05	-1.76
County_Agg53_63	-0.03	0.05	-0.62
County_Agg71_81	-0.02	0.05	-0.33

136. **Step 4 – Calculate average characteristics** - Calculate the quantity of the characteristics for the "typical" property by calculating the mean for each numerical variable and the frequency for each qualitative/dummy variable in the reference period (1Q2008).

Table 40. Characteristics – Coefficients and Average Characteristics for New Dwellings in the Reference Period (1Q2008) and in the Current Period (2Q2008)

CHR	Coef_P2s1	Coef_P1s1	Average
Intcpt	11.8	11.9	
BERHigh	-0.24	-0.28	0.98
Building_Levels1	0.00	0.08	0.05
Central_HeatingElectricity	0.04	-0.14	0.08
Central_HeatingMains_Gas	0.08	0.07	0.54
Central_HeatingOil	-0.07	-0.15	0.35
County_Agg11_14	-0.18	-0.32	0.12
County_Agg22_31	-0.24	-0.18	0.09
County_Agg31_41	0.01	-0.12	0.18
County_Agg42_52	0.01	-0.09	0.17
County_Agg53_63	0.02	-0.03	0.17
County_Agg71_81	0.13	-0.02	0.21
Dwelling_TypeApartment	-0.09	0.00	0.22
Dwelling_TypeDetached	0.10	0.15	0.24
Dwelling_TypeSemi_Detached	0.06	0.14	0.33
Floor_Area	0.00	0.00	122.83
Neighborhood_Affluence	0.01	0.01	23.90
Neighborhood_TypeRural	-0.10	-0.01	0.27

Intcpt - Intercept

Coef_P2s1 – Coefficients of period 2 (2Q2008) in stratum 1

Coef_P1s1 – Coefficients of period 1 (1Q2008) in stratum 1

Average – Mean or frequency of the characteristics in the reference period (1Q2008)

137. **Step 5 - Compile the sub-indices** - The sub-indices compilation is as show in equation (5). The results of the calculations for period 2Q2008 are shown in Table 41 and the result shows a decrease in the prices of dwellings of 5.5 percent since 1Q2008.

Table 41. Characteristics – Compilation of the Sub-Index for New Dwellings in the Current Period (2Q2008)

CHR	Coef_P2s1	Coef_P1s1	Average	
Intcpt	11.8	11.9		11.8 - 11.9
BERHigh	-0.24	-0.28	0.98	(- 0.24 - (- 0.28) * 0.98
Building_Levels1	0.00	0.08	0.05	(0.0 - 0.08) * 0.05
Central_HeatingElectricity	0.04	-0.14	0.08	(0.04 - (- 0.14) * 0.08
Central_HeatingMains_Gas	0.08	0.07	0.54	(0.08 - 0.07) * 0.54
Central_HeatingOil	-0.07	-0.15	0.35	(-0.07 - (- 0.15)) * 0.35
County_Agg11_14	-0.18	-0.32	0.12	(-0.18 - (- 0.32) * 0.12
County_Agg22_31	-0.24	-0.18	0.09	(-0.24 - (- 0.18) * 0.09
County_Agg31_41	0.01	-0.12	0.18	(0.01 - (- 0.12) * 0.18
County_Agg42_52	0.01	-0.09	0.17	(0.01 - (- 0.09) * 0.17
County_Agg53_63	0.02	-0.03	0.17	(0.02 - (- 0.03) * 0.17
County_Agg71_81	0.13	-0.02	0.21	(0.13 - (- 0.02) * 0.21
Dwelling_TypeApartment	-0.09	0.00	0.22	(- 0.09 - 0.00) * 0.22
Dwelling_TypeDetached	0.10	0.15	0.24	(0.10 - 0.15) * 0.24
Dwelling_TypeSemi_Detached	0.06	0.14	0.33	(0.06 - 0.14) * 0.33
Floor_Area	0.00	0.00	122.83	(0.00 - 0.00) * 122.83
Neighborhood_Affluence	0.01	0.01	23.90	(0.01 - 0.01) * 23.90
Neighborhood_TypeRural	-0.10	-0.01	0.27	(-0.10 - (- 0.01) * 0.27
Sub-index New Dwellings				**Exp(sum(…)) * 100 = 94.5**

* User may not obtain the same result due to rounding.

138. Step 6 – Aggregation (Laspeyres formula) – Aggregate the sub-indices (stratum 1 and stratum 2) using the property values as weights as shown in exercise 1.

Table 42. Characteristics – Aggregation of the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Weights
RPPI	100	96.1	94.2	90.0	1
New	100	94.5	91.4	88.5	0.24
Existing	100	96.5	95.1	90.5	0.76

139. Step 7 – Re-reference the sub-indices - The indices initially compiled using the first quarter of 2008 as the reference period are re-referenced such that 2008=100. Table 43 shows a price increase between the first and second quarter of one percent compared to the average of 2008.

Table 43. Characteristics – Re-referencing the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Re-referencing factor (mean of indices)
RPPI	100 / 0.95 = 105.2	96.1 / 0.95 = 101.0	94.2 / 0.95 = 99.1	9.0 / 0.95 = 94.7	(100 + 96.1 + 94.2 + 90.0) = 0.95
New	100 / 0.94 = 106.9	94.5 / 0.94 = 101.0	91.4 / 0.94 = 97.6	88.5 / 0.94 = 94.5	(100 + 94.5 + 91.4 + 88.5) = 0.94
Existing	100 / 0.96 = 104.7	96.5 / 0.96 = 101.0	95.1 / 0.96 = 99.6	90.5 / 0.96 = 94.7	(100 + 96.5 + 95.1 + 90.5) = 0.96

140. The index compilation for the following years is described in the next five steps. The calculations can be replicated using the R procedure named "Stratification_production.r." The average characteristics and the coefficients of the base period can be obtained by using the R procedure named "base_characteristics.r" for each year using data from the fourth quarter. These should be updated every year using the last quarter of the previous year

141. Steps 1 to 3 are the same as for the first year.

142. Step 4 is calculated with the R procedure "base_characteristics.r."

143. Step 5 and 6 are again the same as previously noted above.

144. Step 7 - Chain the indices. To express the indices in terms of the reference year they need to be chained. Chain the index by multiplying the index for the current period by the chained index of the last period of the previous year as shown in Table 44 and Table 45.

Table 44. Characteristics – Chaining the Indices for 2009

	4Q 2008	1Q2009		2Q2009		3Q2009		4Q2009	
	2008= 100	4Q2008= 100	2008= 100	4Q2008= 100	2008= 100	4Q2008= 100	2008= 100	4Q2008= 100	2008= 100
RPPI	94.7	94.2	89.2	90.6	85.7	85.8	81.2	80.8	76.5
New	94.5	95.0	89.8	93.5	88.4	83.8	79.2	79.5	75.1
Existing	94.7	94.0	89.0	89.6	84.9	86.4	81.9	81.2	76.9

Table 45. Characteristics – Chaining the Indices for 2010

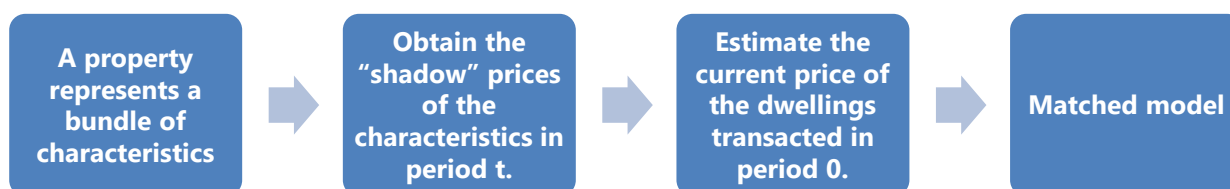|	4Q 2009	1Q2010		2Q2010		3Q2010		4Q2010	
	2008= 100	4Q2009 =100	2008= 100	4Q2009 =100	2008= 100	4Q2009 =100	2008= 100	4Q2009 =100	2008= 100
RPPI	76.5	94.7	72.5	93.0	71.2	93.3	71.4	92.3	70.6
New	75.1	100.9	75.8	99.3	74.6	93.6	70.3	89.8	67.5
Existing	76.9	93.8	72.2	92.1	70.8	93.3	71.8	92.6	71.3

H. Method 4: Imputations Hedonic Method[10]

Reference: Handbook on Residential Property Price Indices, pages 54-56

145. Compilers may also want to consider using the imputations hedonic method. The imputations hedonic method imputes a price for each property for each period (regardless as of whether a transaction occurred) and calculates an index using the imputed prices.

146. This method requires detailed information on property characteristics and a large number of transactions. It is generally appropriate if geospatial data (coordinates) are used. If the log-linear form is used, the characteristics and imputation hedonic methods are equivalent, except in cases where geospatial data are used.

Figure 7. The Mechanics of the Imputations Hedonic Method

| A property represents a bundle of characteristics | → | Obtain the "shadow" prices of the characteristics in period t. | → | Estimate the current price of the dwellings transacted in period 0. | → | Matched model |

147. The log-linear specification for each stratum is used:

$$\ln p_n^t = \beta_0 + \sum_{k=1}^{K} \beta_k z_{nk}^t + \varepsilon_n^t \tag{6}$$

148. Separate regressions for reference period (0) and the current period (t) are used to estimate the "shadow" prices, or the estimated coefficients ($\hat{\beta}$), for each quarter in a stratum. After exponentiating, the imputed prices for the reference period (0) are:

$$\hat{p}_n^0(0) = \exp\left(\widehat{\beta}_0^0\right) \exp\left[\sum_{k=1}^{K} \widehat{\beta}_k^0 z_{nk}^0\right] \tag{7}$$

149. The imputed prices for the current period (t) for the dwellings sold in the reference period (0) are shown in equation (8).

$$\hat{p}_n^t(0) = \exp\left(\widehat{\beta}_0^t\right) \exp\left[\sum_{k=1}^{K} \widehat{\beta}_k^t z_{nk}^0\right] \tag{8}$$

150. The index is then calculated, by taking the ratio of the geometric mean of the imputed prices for the reference period dwellings in period t to the geometric average of the imputed prices for the same dwellings in reference period 0:

$$I_t = \frac{\prod_{n \in S(0)} (\hat{p}_n^t(0))^{1/N(0)}}{\prod_{n \in S(0)} (\hat{p}_n^0)^{1/N(0)}} \tag{9}$$

---

[10] Also called hedonic re-pricing method.

151.   Note that the observed prices from the reference period (0) are replaced by the imputed values. This is because the estimates resulting from the omitted variables offset each other to some extent in both periods.

152.   As with the other methods it is advisable to update the reference period dwellings each year. For example, for a quarterly index, the dwellings from the fourth quarter in year t can be used as the reference period dwellings for the index calculation in year t+1.

153.   To help compilers become familiar with the process to calculate an RPPI using the imputations hedonic method an exercise has been developed to accompany this *Guide*. The exercise (Exercise 5) uses a synthetic data set and a set of procedures written in R. The following step-by-step instructions are intended to walk the reader through the exercise.

Exercise 5: Step-by-step Calculations for a Quarterly Index Using the Imputations Hedonic Method

154.   The index compilation for the first year is presented in the following eight steps. The calculations can be replicated using the R procedure named "imputations_starting.r" and the synthetic set of data accompanying this *Guide*

155.   Step 1 - Obtain data.

156.   Step 2 - Build the strata – Construct the stratum as per exercise 1. Perform the following steps for each stratum.

157.   Step 3 - Calculate the "shadow" prices - Run a regression for each period with the price logarithm as the dependent variable and all other significant variables as the explanatory variables. The coefficients from the regression are the "shadow prices" of the variables.

Table 46. Imputations – Coefficients

	1Q2008	2Q2008	3Q2008	4Q2008
Intcpt	11.90	11.80	11.60	11.72
BERHigh	-0.28	-0.24	-0.25	-0.30
Building_Levels1	0.08	0.00	0.09	0.02
Central_HeatingElectricity	-0.14	0.04	0.01	-0.12
Central_HeatingMains_Gas	0.07	0.08	0.10	0.08
Central_HeatingOil	-0.15	-0.07	-0.05	-0.07
County_Agg11_14	-0.32	-0.18	-0.12	-0.18
County_Agg22_31	-0.18	-0.24	-0.09	-0.13
County_Agg31_41	-0.12	0.01	0.04	0.01
County_Agg42_52	-0.09	0.01	0.11	-0.02
County_Agg53_63	-0.03	0.02	0.00	-0.06
County_Agg71_81	-0.02	0.13	0.12	-0.04
Dwelling_TypeApartment	0.00	-0.09	-0.06	0.03
Dwelling_TypeDetached	0.15	0.10	0.12	0.05
Dwelling_TypeSemi_Detached	0.14	0.06	0.02	-0.03
Floor_Area	0.00	0.00	0.00	0.00
Neighborhood_Affluence	0.01	0.01	0.01	0.01
Neighborhood_TypeRural	-0.01	-0.10	-0.03	-0.04

158. **Step 4 – Impute the prices** - For each property impute a price for each period (including the reference period) using the coefficients from the regressions.

Table 47. Imputations – Example of Price Imputations of Three Dwellings from New Dwellings of Period 1Q2008

	Estimate		ID 6		ID 18		ID 23
Intcpt	11.90		11.90		11.90		11.90
BERHigh	-0.28	1	-0.28	1	-0.28	1	-0.28
Building_Levels1	0.08	0	0.00	1	0.08	1	0.08
Central_HeatingElectricity	-0.14	0	0.00	0	0.00	0	0.00
Central_HeatingMains_Gas	0.07	1	0.07	0	0.00	1	0.07
Central_HeatingOil	-0.15	0	0.00	1	-0.15	0	0.00
County_Agg11_14	-0.32	0	0.00	0	0.00	0	0.00
County_Agg22_31	-0.18	0	0.00	0	0.00	0	0.00
County_Agg31_41	-0.12	0	0.00	0	0.00	0	0.00
County_Agg42_52	-0.09	0	0.00	0	0.00	0	0.00
County_Agg53_63	-0.03	1	-0.03	1	-0.03	0	0.00
County_Agg71_81	-0.02	0	0.00	0	0.00	1	-0.02
Dwelling_TypeApartment	0.00	0	0.00	0	0.00	0	0.00
Dwelling_TypeDetached	0.15	0	0.00	1	0.15	0	0.00
Dwelling_TypeSemi_Detached	0.14	0	0.00	0	0.00	0	0.00
Floor_Area	0.00	94	0.41	152	0.66	86	0.37
Neighborhood_Affluence	0.01	35	0.40	18	0.20	25	0.29
Neighborhood_TypeRural	-0.01	0	0.00	1	-0.01	0	0.00
log imputed price			12.46		12.52		12.41
imputed price			257699.00		274896.24		246011.28

159. **Step 5 - Aggregate the estimated prices** - The imputed prices are then aggregated by taking the geometric mean of the imputed prices for the properties in each stratum for each period.

Table 48. Imputations – Jevons of the Imputed Prices for the Year 2008

Jevons	1Q2008	2Q2008	3Q2008	4Q2008
New	233573	220722	213389	206634
Existing	239609	231316	227946	216885

160. **Step 6 - Calculate the sub-indices** – The stratum index is the ratio of the stratum level imputed price in period t to the imputed price in period 0 as follows:

$$I_t = \text{Geomean}\ (\hat{p}_n^t(0))\ /\ \text{Geomean}\ (\hat{p}_n^0(0)) \qquad (10)$$

Table 49. Imputations – Compilation of the Sub-indices for the Year 2008

Sub-Indices	1Q2008	2Q2008	3Q2008	4Q2008
New	100	220722 / 233573 * 100 = 94.5	213389 / 233573 * 100 = 91.4	206634 / 233573 * 100 = 92.9
Existing	100	231316 / 239609 * 100 = 96.5	227946 / 239609 * 100 = 95.1	216885 / 239609 * 100 = 90.5

161. **Step 7 – Aggregation (Laspeyres)** - The aggregation of the sub-indices is a weighted average using the property values as weights as shown in exercise 1.

Table 50. Imputations – Aggregation of the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Weights
RPPI	100	96.1	94.2	91.1	1
New	100	94.5	91.4	92.9	0.24
Existing	100	96.5	95.1	90.5	0.76

162. **Step 8 – Re-reference the sub-indices** - The indices initially compiled in the reference period 1Q2008 need to be referenced such that 2008=100. This is done by first taking the mean of the quarterly indices and then dividing the quarterly indices by the mean.

Table 51. Imputations - Re-Referencing the Sub-Indices

	1Q2008	2Q2008	3Q2008	4Q2008	Referencing Factor (Mean of Indices)
RPPI	100 / 0.95 = 104.9	96.1 / 0.95 = 100.7	94.2 / 0.95 = 98.8	91.1 / 0.95 = 95.5	(100 + 96.1 + 94.2 + 91.1) / 4 = 0.95
New	100 / 0.95 = 105.6	94.5 / 0.95 = 99.8	91.4 / 0.95 = 96.5	92.9 / 0.95 = 98.1	(100 + 94.5 + 91.4 + 92.9) / 4 = 0.95
Existing	100 / 0.96 = 104.7	96.5 / 0.96 = 101.0	95.1 / 0.96 = 99.6	90.5 / 0.96 = 94.7	(100 + 96.5 + 95.1 + 90.5) / 4 = 0.96

163. The index compilation for the following years (2009 and 2010) is presented in the following five steps. The calculations can be replicated using the R procedure named "imputations_production.r". The base price used to compile the index for 2009 and 2010 can be obtained with the R procedure named "base price.r".

164. **Steps 1 to 5 are the same as for the first year**.

165. **Step 6 - Calculate the sub-indices** - The sub-indices are compiled as the ratio of the geometric mean of the current period imputed price to the geometric mean of the base period imputed price (last quarter of previous year).

Table 52. Imputations – Compilation of the Sub-Indices for 1Q2009

	Jevons 4Q2008	Jevons 1Q2009	Sub-indices 1Q2009
New	206634	199271.3	199271.3 / 206634 * 100 = 95.0
Existing	216885	195873.1	195873.1 / 216885 * 100 = 94.0

166.   Step 7 – Aggregation (Laspeyres) - The aggregation of the sub-indices uses the weights as calculated in exercise 1.

167.   Step 8 - Chain the indices - Chain the index by multiplying the index of the current period by the chained index of the last period of the previous year as shown in Table 53 and Table 54.

Table 53. Imputations – Chaining the Indices for 2009

	4Q2008	1Q2009		2Q2009		3Q2009		4Q2009	
	2008=100	4Q2008 =100	2008= 100	4Q2008 =100	2008= 100	4Q2008 =100	2008= 100	4Q2008 =100	2008= 100
RPPI	100.5	100.8	101.2	102.5	103.0	102.7	103.2	99.7	100.2
New	100.0	101.3	101.3	102.8	102.8	102.6	102.7	101.2	101.2
Existing	101.3	99.8	101.1	101.9	103.2	102.7	104.1	97.1	98.4

Table 54. Imputations – Chaining the Indices for 2010

|	4Q2008	1Q2009		2Q2009		3Q2009		4Q2009	
	2008= 100	4Q2008 =100	2008= 100	4Q2008 =100	2008= 100	4Q2008 =100	2008= 100	4Q2008 =100	2008= 100
RPPI	95.5	94.2	89.2	90.6	85.7	85.8	81.2	80.8	76.5
New	98.1	95.0	89.8	93.5	88.4	83.8	79.2	79.4	75.1
Existing	94.7	94.0	89.0	89.6	84.9	86.4	81.9	81.2	76.9

I. RPPI Results

168.   Table 55 and Figure 8 shows the RPPI compiled using the four methods described in this *Guide*. The characteristics and the imputations methods provide similar results since they are variants of the same method. As expected, the time dummy provides less volatile results since it uses a rolling window of four quarters which helps smooth the index. In contrast, the RPPI calculated with a simple stratification method using a median price is highly volatile due to the lack of proper quality adjustment.

Table 55. RPPI Index Compiled with Four Methods

RPPI	1Q2008	2Q2008	3Q2008	4Q2008	1Q2009	2Q2009	3Q2009	4Q2009	1Q2010	2Q2010	3Q2010	4Q2010	1Q2011	2Q2011	3Q2011	4Q2011
Stratification	105.0	103.3	98.3	93.4	88.2	86.3	81.3	76.9	70.7	71.5	71.7	70.0	60.7	64.6	67.4	67.9
Time dummy	105.2	101.0	99.1	94.7	89.2	85.7	80.8	76.1	71.7	70.8	71.0	70.1	66.5	68.0	70.0	70.5
Characteristics	105.2	101.0	99.1	94.7	89.2	85.7	81.2	76.5	72.5	71.2	71.4	70.6	67.0	68.8	70.7	71.6
Imputations	105.2	101.0	99.1	94.7	89.2	85.7	81.2	76.5	72.5	71.2	71.4	70.6	67.0	68.8	70.7	71.6

RPPI	1Q2012	2Q2012	3Q2012	4Q2012	1Q2013	2Q2013	3Q2013	4Q2013	1Q2014	2Q2014	3Q2014	4Q2014	1Q2015	2Q2015	3Q2015	4Q2015
Stratification	64.1	67.6	72.7	70.1	70.3	75.7	80.5	80.4	80.8	81.6	87.7	86.9	91.2	89.6	98.2	97.5
Time dummy	70.0	74.3	77.5	78.0	78.6	83.0	85.7	86.1	87.5	89.7	93.1	94.3	97.0	97.4	104.4	105.6
Characteristics	71.5	76.2	79.0	79.4	80.0	84.4	87.3	87.4	88.9	91.1	94.5	95.9	98.7	99.0	106.1	107.7
Imputations	71.5	76.2	79.0	79.4	80.0	84.4	87.3	87.4	88.9	91.1	94.5	95.9	98.7	99.0	106.1	107.7

Figure 8. RPPI Index Compiled with Four Methods
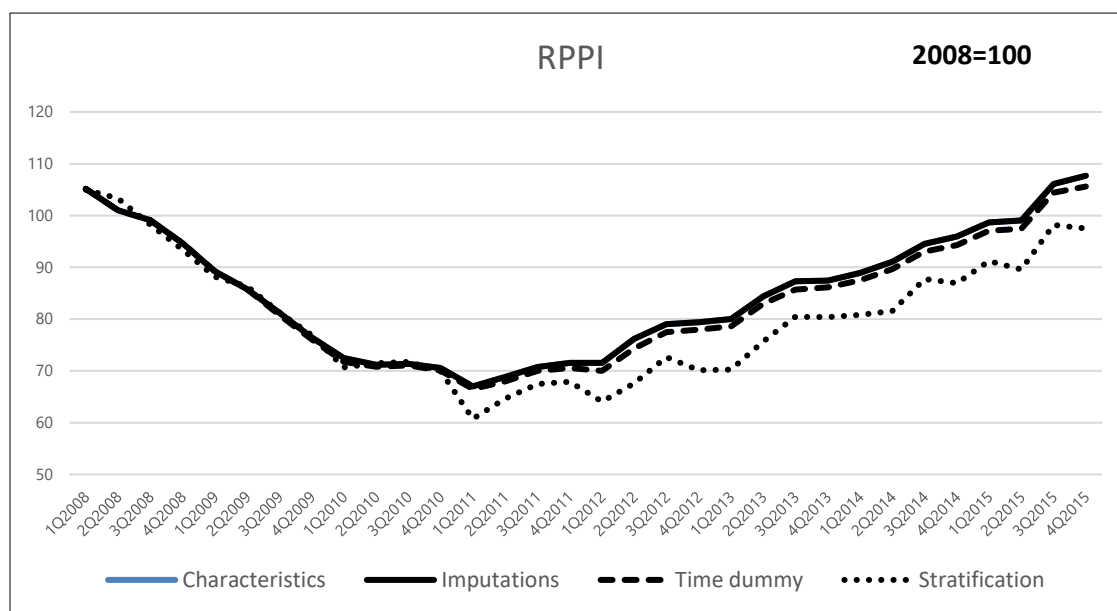
DISSEMINATION AND OTHER INDICATORS

169.   This *Guide* has provided compilers with a series of step-by-step exercises that demonstrate the different ways an RPPI can be computed. The choice of method depends largely on the availability of data and the needs of the users. The methods that utilize hedonic regressions are generally considered superior and provide users with higher quality estimates. If compilers are unable to acquire the necessary data to use a hedonic-based method, other methods such as median with stratification may be appropriate.

170.   Before disseminating official results, compilers may consider developing experimental indices for at least one year. The experimental indices can be vetted by various users in terms of their accuracy, relevance and timeliness. The compiler can then apply the feedback they receive to refine both the methods and business processes used to construct the index. The development of experimental estimates also allows the compiler to test their overall business process and make any necessary adjustments prior to the official release of the data. Once the methodology and business process have been finalized, the index should be compiled for several periods before disseminating the official results.

171.   RPPIs are an important source of information used by policy makers to better understand financial stability as well as socio-economic issues such as housing affordability. Given the extensive use of the indices it is important that compilers develop an effective dissemination strategy. The indices should be made public with a media release. Compilers may also want to organize a meeting or seminar with the main stakeholders at the time of the release where they can explain the index and how it can (and should not) be interpreted. Compilers may want to consider the use of social media and videos to ensure awareness of the index is as broad-based as possible.

172.   The release should also be accompanied by a technical note on sources and methods clearly explaining the methodology. The technical note should be a short document that is clear, easy to read, and includes some of the following information:

- the aim of the index;
- the coverage;
- the stratification;
- the weighting system;
- the reference periods of the weights and the indices; and
- the method used.

173.   The technical note is the main document stakeholders will consult when assessing the credibility, quality, and usability of the index. The presentation of the technical note should follow a standard format such as the IMF factsheet for the Data Quality Assessment Framework, the United Nations Generic National Quality Assurance Framework or the template recommended by the Second Phase of the G20 Data Gaps Initiative as in Annex II.

174. A well-articulated technical note or sources and methods guide is especially important in the case of an RPPI due to the diversity of methods that can be used for its compilation. International comparability can be deeply affected by the uses of different methodologies. For this reason, it is crucial to provide information on how the index is compiled. Over time the data sources and methods used to compile the index may change. The compiler should ensure users are informed of the developments. RPPIs are seen as a leading indicator that can help inform users about the business cycle. Compilers should target releasing the data within 90 days after the reference period for quarterly data. Compilers should make the data available in machine readable form (e.g., .csv) and publish the data according to an advance release calendar.

175. A general recommendation is to publish indices with two decimals and rates with one decimal.

176. The publication of a time-series graph with the RPPI and its sub-indices is useful. The graph should clearly indicate the reference period (base) as shown in Figure 9 where the reference period is 2008. Figure 9 shows an example where the total RPPI and sub-indices for new and existing dwellings all decline from 2008 through l 2011, then recover beginning in 2012.

Figure 9. RPPI and Sub-Indices
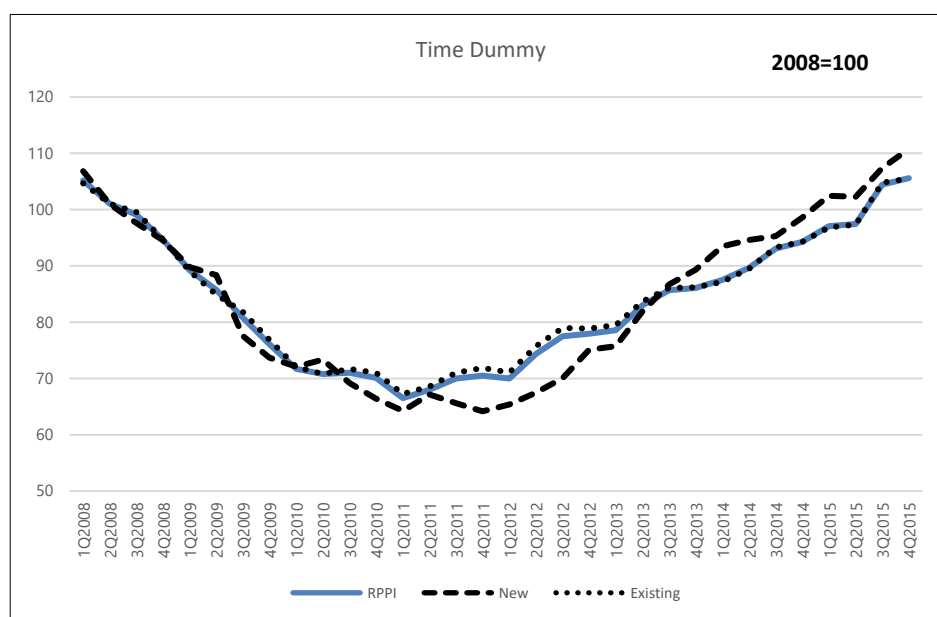
177. The trend of the contribution of each stratum to the annual change (YoY) in the total RPPI can be better displayed in a graph such as Figure 10. As expected, due to its larger weight, the index for existing dwellings contributes most of the change in the overall RPPI.

Figure 10. YoY and Contributions
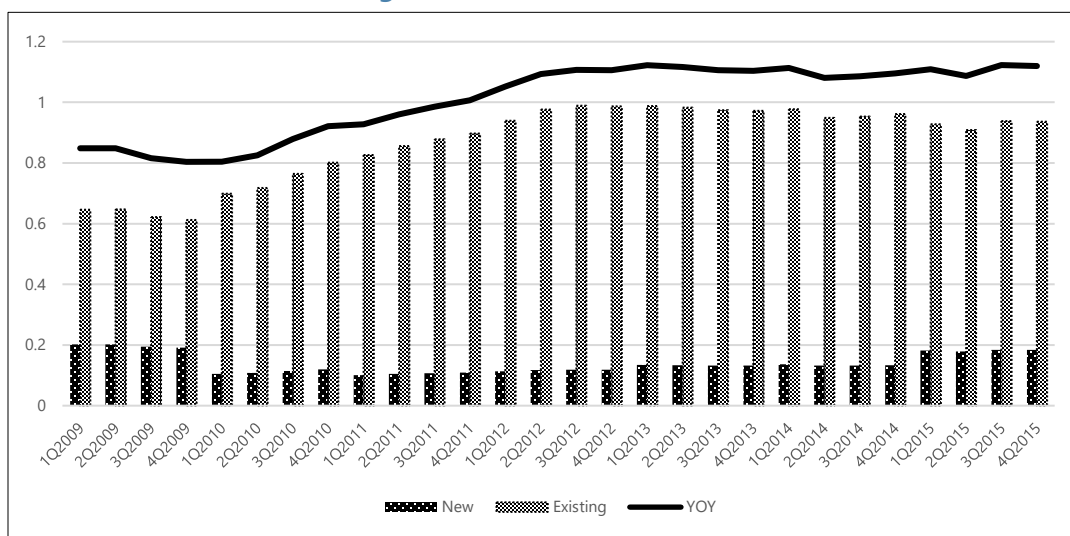
178. The trend of the RPPI can be compared with the development of the number of dwellings sold in each period. Figure 11 shows that for both strata (sub-indices) the number of sales follow the same trend as the RPPI, from 3Q2010 but not before. However, there is a seasonal pattern where sales peak at the end of each year.

Figure 11. RPPI and Number of Dwellings Sold
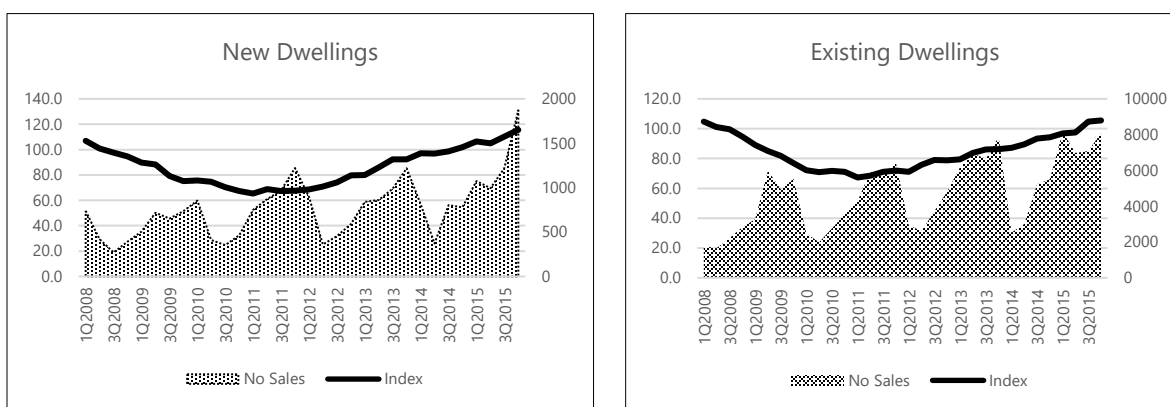
179. The change in the prices of dwellings may or not follow the general inflation trend. Benchmarking the RPPI with the CPI[11] may provide some information of how in line are the house price indices with the general inflation. In Figure 12 we can see a higher volatile in house prices than in consumer prices. While consumer prices after a short fall in 2008 remain rather stable until the end of 2010 and followed by a slow increase trend, the house prices register a sharp decrease until 2001 followed by a high increase.

---

[11] The CPI values shown do not relate to any specific country. These are random values to illustrate the idea.

Figure 12. RPPI and CPI

180. Comparing the trend in a national RPPI with the RPPIs of neighboring countries can also provide useful information for policy making. However, this analysis must take into account whether different methodologies are used to compile the RPPI in the comparator countries.

181. Finally, spreadsheets with the RPPI series should also be available for download from the website.

182. More and more statistical agencies are experimenting with various ways to visually present RPPIs. RPPIs lend themselves to visual presentation as they are multi-dimensional with respect to time, geography, dwelling type, etc.).

183. The FT visual vocabulary provides guidance on how to create easily readable graphs for different types of information. The image bellow shows some examples of how to build graphs for indicators measuring change over time.

Figure 13. The Financial Times Vocabulary

Deviation   Correlation Change v Time   Ranking   Distribution Part to whole Magnitude   Spatial   Flow
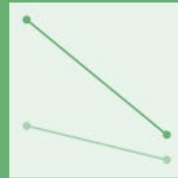
Change v Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries: Choosing the correct time period is important to provide suitable context for the reader

Examples of use
Share price movements, economic time series

Chart types

line	column-timeline	column-line-timeline	stock-price	slope	area
The standard way to show a changing time series. If data are irregular, consider markers to represent data points	Columns work well for showing change over time - but usually best with only one series of data at a time	A good way of showing the relationship over time between an amount (columns) and a rate (line)	Usually focused on day-to-day activity, these charts show opening/closing and hi/low points of each day	Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story	Use with care. These are good at showing changes to total, but seeing change in components can be very difficult.

184.   Compilers can also experiment with the use of maps and other less traditional visualizations to ensure their data is as impactful as possible. For example, Figure 14 shows the cities in the United States with the highest concentration of dwellings over a certain value.

Figure 14. US Cities with the Highest Share of Million-Dollar Dwellings
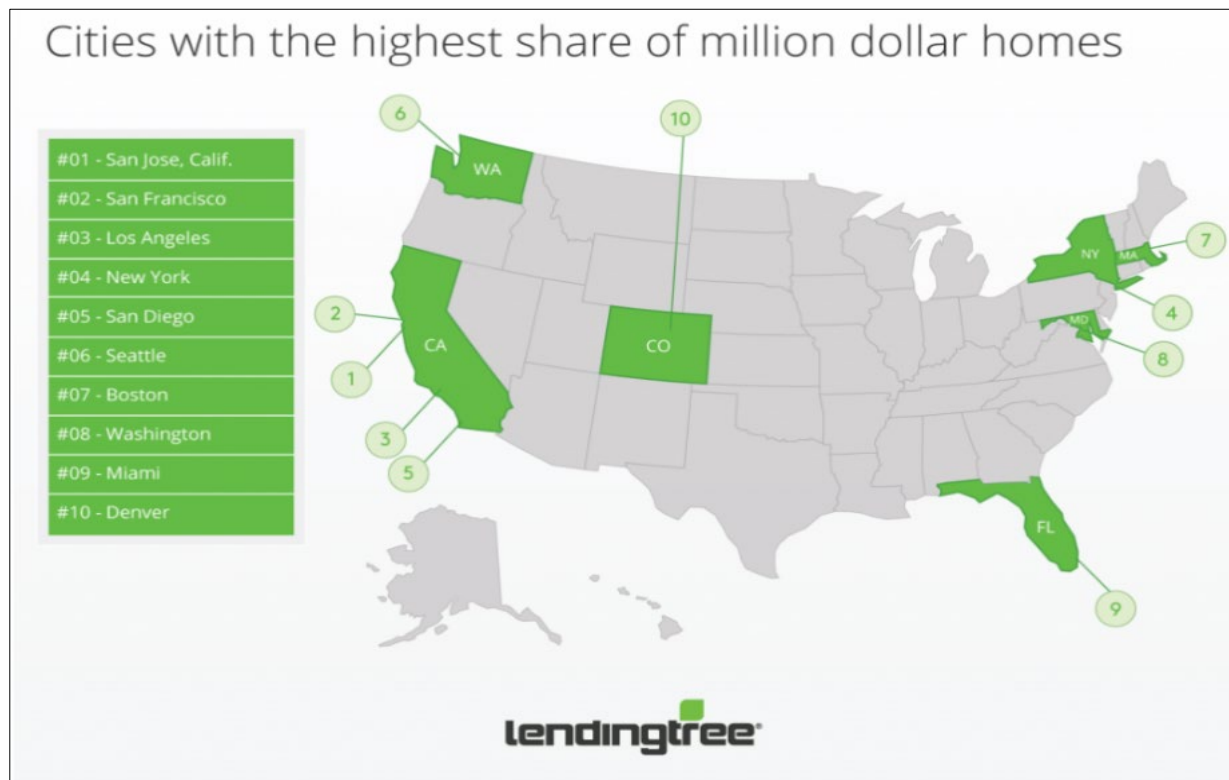

Source: Lendingtree.

185.   Many statistical agencies are releasing RPPIs within a broader set of residential property information. RPPIs and the data that underpin their calculation can be used to examine issues such as housing affordability. Consider the examples below which use the price levels in domestic currency as an input to examine housing affordability in Sweden as well as across a select number of other countries. These indicators can be calculated using the RPPI input data in conjunction with income data. Figure 15 shows the ratio of house prices to income for cities in Sweden while Figure 16 compares house prices to income for selected countries. These indices were constructed using the same data that are used to compile RPPIs.

Figure 15. Housing Affordability in Sweden Sample of Cities
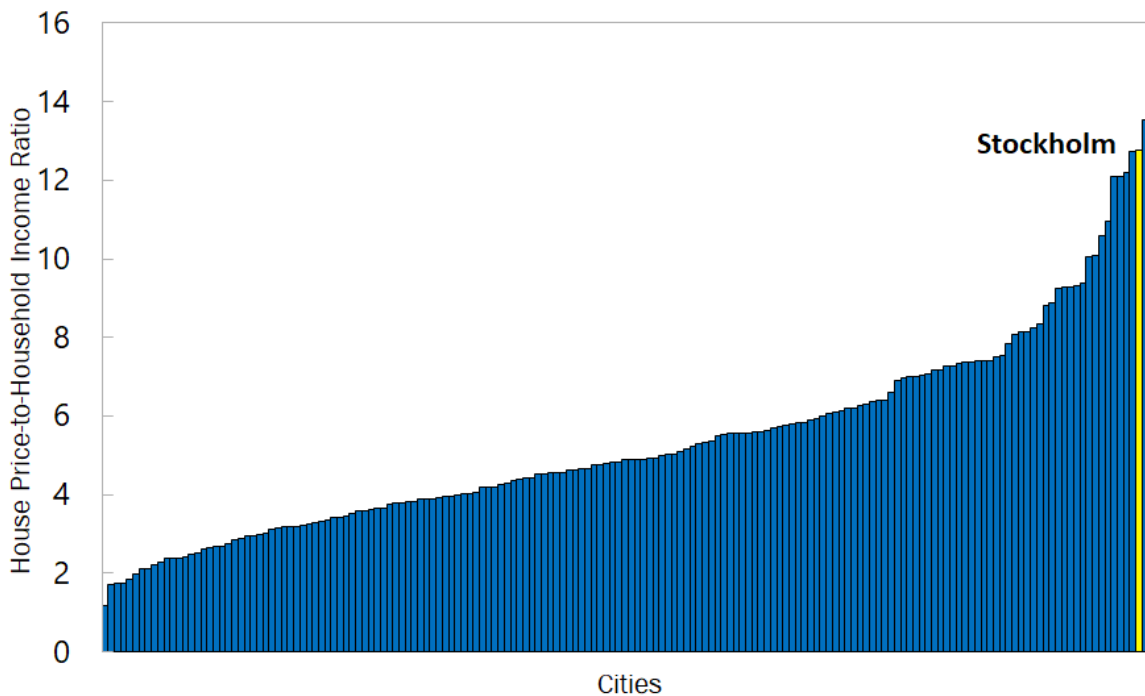
Sources: The Land and Housing Survey in the UN Sample of Cities (2016), Statistics Sweden, and Fund Staff calculations.

**Figure 16. House Price-to-Income Ratio Across Countries (As of 2016Q4)**
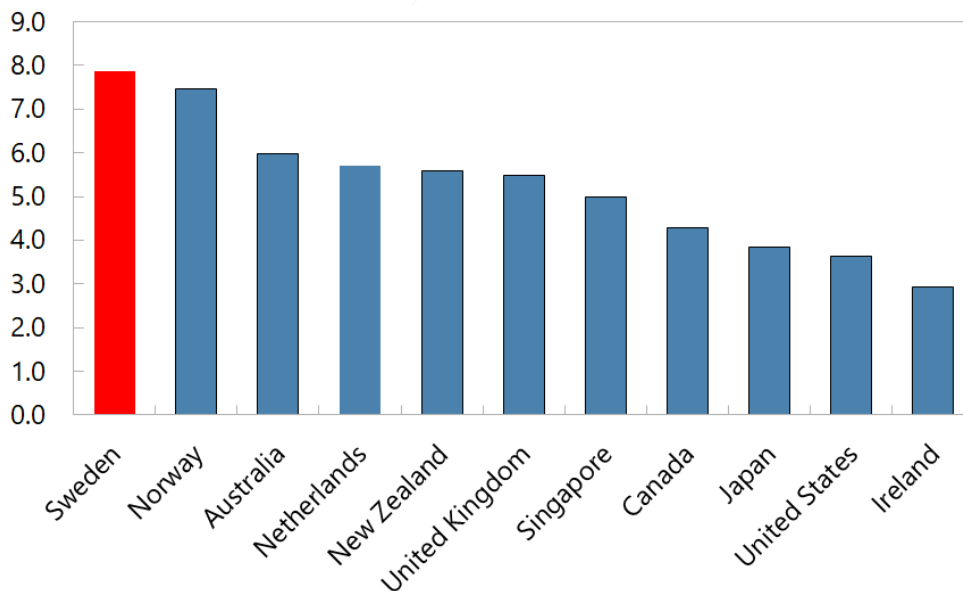
Sources: Demographia International Housing Affordability Survey, OECD, Statistics Netherlands, Statistics Norway, Statistics, Sweden, and Fund staff calculations.

186. The Second Phase of the G-20 Data Gaps Initiative recommends statistical agencies release indicators on the number of dwelling sold. Indicators on sales generally appeal to a broader

audience then a price index. Figure 17 provides an example of how data on housing sales can be disseminated to the general public.

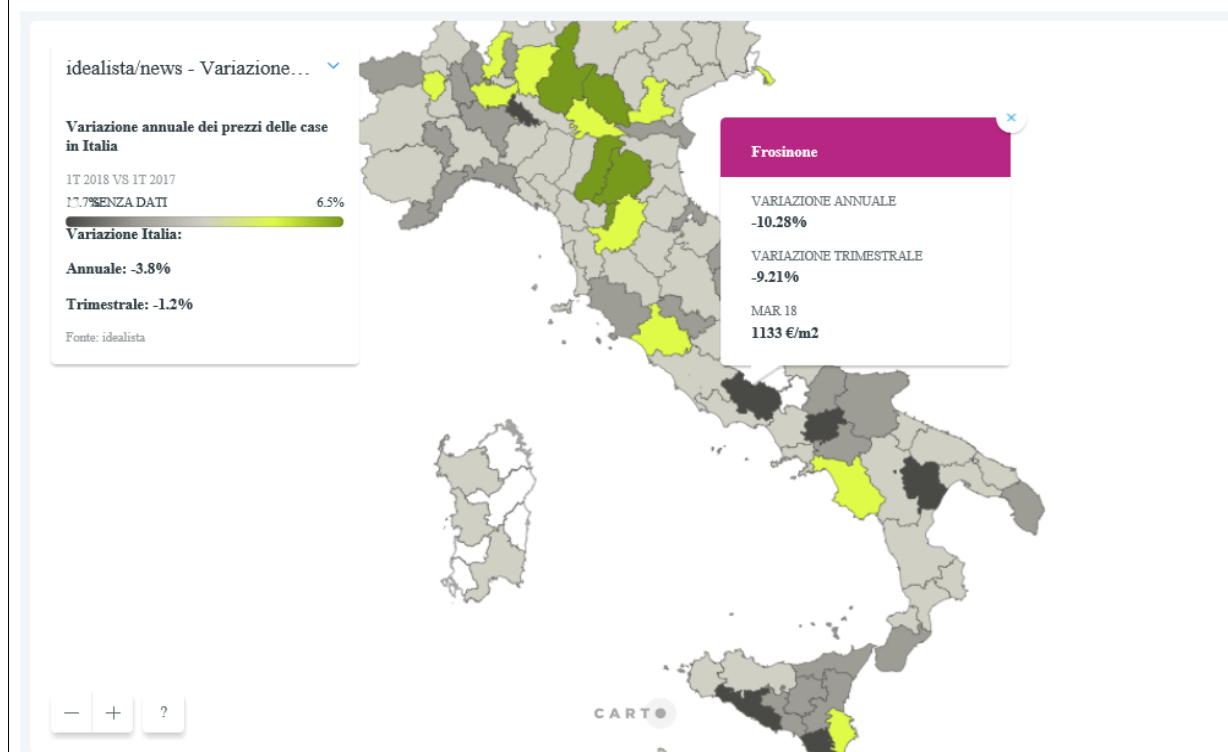**Figure 17. Residential Property Sales Variation**
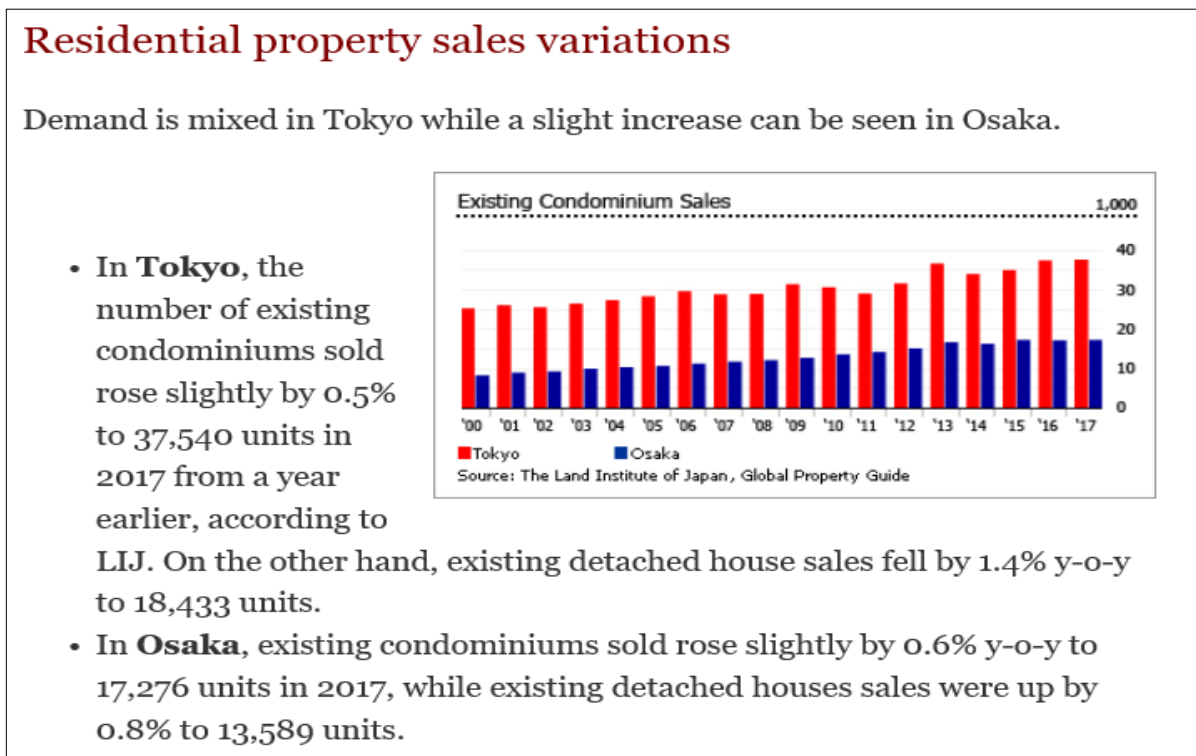
Source: Global Property Guide

187. Regional price development and prices per square meter/foot can be of interest for the public, especially when trends differ from region to region. The interactive graphic below provides a wealth of housing related information for the different regions in Italy.

Figure 18. Heat Map with Price Development Per Square Meter in Italy

Prezzi della case ancora in calo nel 2018: mappa variazione annuale e trimestrale, valori al m2

idealista/news - Variazione...

Variazione annuale dei prezzi delle case in Italia

1T 2018 VS 1T 2017

17.7% SENZA DATI   6.5%

Variazione Italia:

Annuale: -3.8%

Trimestrale: -1.2%

Fonte: idealista

Frosinone

VARIAZIONE ANNUALE
-10.28%

VARIAZIONE TRIMESTRALE
-9.21%

MAR 18
1133 €/m2

Source: idealista.it.

188.   Dissemination is often an afterthought in the compilation of RPPIs. To ensure the results are as impactful as possible, compilers are encouraged to develop a dissemination strategy prior to the development of the index. The strategy should consider user needs, the latest technical advances in the area of data visualization as well as the diverse set of uses of the information. Ideally the dissemination of the index will be part of a larger residential property framework that provides a wealth of information on housing prices, sales, stocks, quality, and affordability.

REFERENCES

Brand, Manuel; Vermeulen, Corinne; Fischbach, Davide; Carpy, Yves (2009). *Swiss residential property price index*

EUROSTAT, ILO, IMF, OECD, UNECE, the World Bank (2013). *Handbook on Residential Property Prices Indices (RPPIs).*

EUROSTAT (2009). *Technical Manual on Owner-Occupied Housing and House Price Indices.*

Geng, Nan (2010). *Fundamental Drivers of House Prices in Advanced Economies.* IMF Working Paper.

Linz, Stefan; Behrmann, Timm (2004). *Using Hedonic Pricing for the German House Price Index.*
IMF the *CPI Manual: Concepts and Methods,* forthcoming.

Lyons, Ronan C. (2019). *Can list prices accurately capture housing price trends? Insights from extreme markets conditions.* Finance Research Letters, 30, 228-232.

Marola, Bogdan; Santos, Daniel; Evangelista, Rui and Guerreiro, Vanda. (2012). *Working Towards a Comparable House Price Inflation Measure in Europe.* UNECE Meeting of the Group of Experts on CPI.

O'Hanlon, Niall (2011). *Constructing a National House Price Index for Ireland.* Journal of the Statistical and Social Inquiry Society of Ireland, 40, 167-192.

Ottawa Group - International Working Group on Price Indices Prud'Homme, Marc and Erdur, Serra (2007).

Rechnio, Renata (2008) *The Polish experience in developing House Price Index.*

Silver, Mick (2008). *How to better measure hedonic residential property price indices.* IMF Working Paper.

UNECE Meeting of the Group of Experts on Consumer Price Indices

Annex I. Relation Between the Methods and the Data Available

No. of Prices (Observations)

Method 1
Average

Method 2
Stratification

Method 4
Characteristics

Method 5
Imputations

Method 3
Rolling window

e.g., town

e.g., town, size,
garage, elevator, ...

geo-spatial

No. of Characteristics
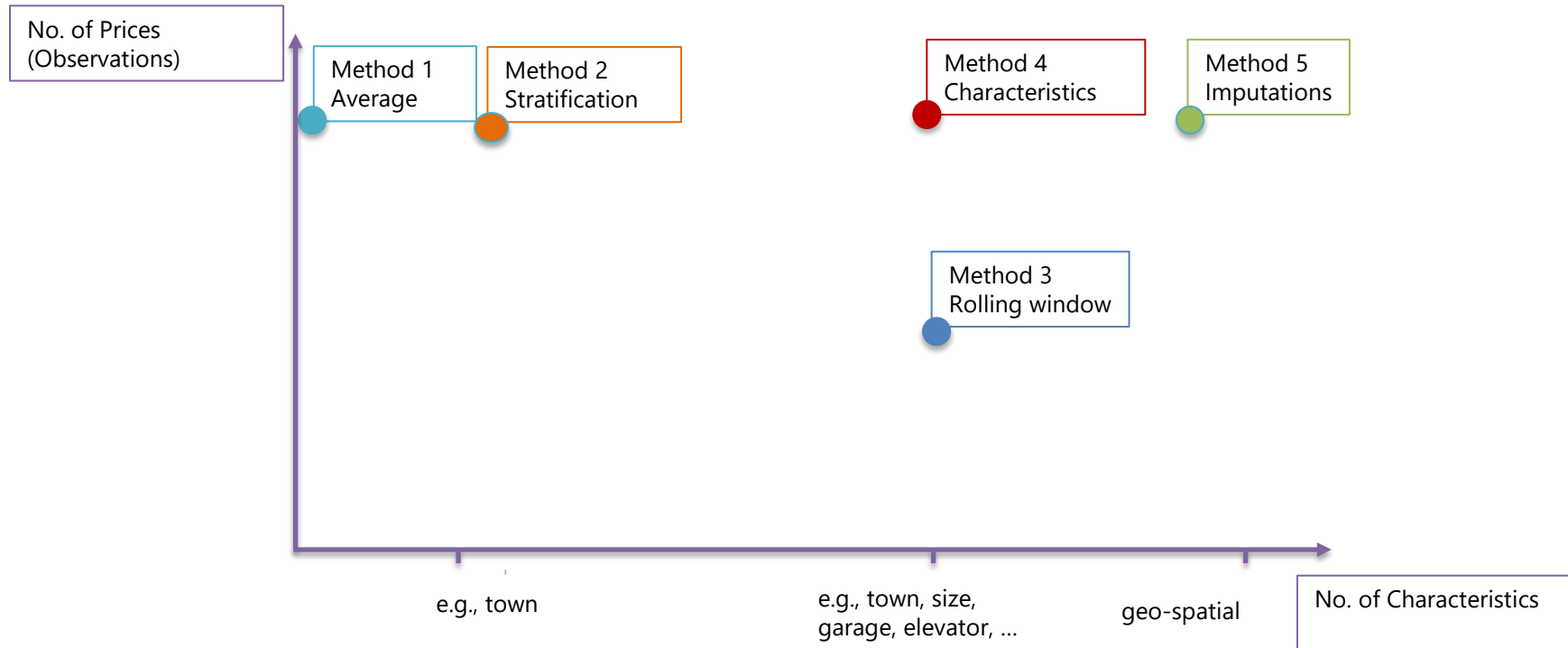
Annex II. G20 Data Gaps Initiative Template on Housing Price Statistics

Residential Property Price Indices (RPPIs)

Tier 1 indicators in green
Tier 2 indicators in orange

Categories	Breakdown		Weighting	Vintage		Type of Dwellings		Geographical Coverage	Frequency (preferred)	Starting Period - Minimum Required	Measures
	Full name of the series	Short Name of the series		New	Existing	Single-unit	Multi-unit				
I.RPPIs	I.1.1. a. RPPI - New and existing dwellings for all types of dwellings (single- and multi-unit[1]) for the whole country	RPPI Total - Whole country	Sales and/or Stock based RPPI[2]	×	×	×	×	Whole country	At least quarterly	2005 Q1	Index - National reference period
	I.1.1. b. RPPI - New and existing dwellings for all types of dwellings (single- and multi-unit[1]) for the capital city[3]	RPPI Total - Capital city[3]	Sales and/or Stock based RPPI[2]	×	×	×	×	Capital City	At least quarterly	2005 Q1	Index - National reference period
	I.1.1. c. RPPI - New and existing dwellings for all types of dwellings (single- and multi-unit[1]) for urban area[4]	RPPI Total - Urban areas[4]	Sales and/or Stock based RPPI[2]	×	×	×	×	Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period
	I.1. 2. a RPPI - New and existing dwellings for single-unit dwellings[1] for the whole country	RPPI Total - Single-unit dwellings[1] - Whole country	Sales and/or Stock based RPPI[2]	×	×	×		Whole country	At least quarterly	2005 Q1	Index - National reference period
	I.1. 2. b RPPI - New and existing dwellings for single-unit dwellings[1] for the capital city[3]	RPPI Total - single-unit dwellings[1] - Capital city[3]	Sales and/or Stock based RPPI[2]	×	×	×		Capital City	At least quarterly	2005 Q1	Index - National reference period
	I.1. 2. c RPPI - New and existing dwellings for	RPPI Total - single-unit	Sales and/or Stock	×	×	×		Urban areas[3]	At least quarterly	2005 Q1	Index - National

single-unit dwellings[1] for urban areas[4]	dwellings[1] - Urban areas[4]	based RPPI[2]								reference period
I.1. 3. a. RPPI - New and existing dwellings for multi-unit dwellings for the whole country	**RPPI Total - Multi-unit dwellings - Whole country**	**Sales and/or Stock based RPPI[2]**	×	×		×	**Whole country**	**At least quarterly**	**2005 Q1**	**Index - National reference period**
I.1. 3. b. RPPI - New and existing dwellings for multi-unit dwellings[1] for the capital city[3]	RPPI Total - Multi-unit dwellings[1] - Capital city[3]	Sales and/or Stock based RPPI[2]	×	×		×	Capital City	At least quarterly	2005 Q1	Index - National reference period
I.1. 3. c. RPPI - New and existing dwellings for multi-unit dwellings[1] for urban areas[4]	RPPI Total - Multi-unit dwellings[1] - Urban areas[4]	Sales and/or Stock based RPPI[2]	×	×		×	Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period
I.2.1.a. RPPI - New dwellings for all types of dwellings (single- and multi-unit[1]) for the whole country	**RPPI New - Whole country**	**Sales and/or Stock based RPPI[2]**	×		×	×	**Whole country**	**At least quarterly**	**2005 Q1**	**Index - National reference period**
I.2.1.b. RPPI - New dwellings for all types of dwellings (single- and multi-unit[1]) for the capital city[3]	RPPI New - Capital city[3]	Sales and/or Stock based RPPI[2]	×		×	×	Capital City	At least quarterly	2005 Q1	Index - National reference period
I.2.1.c. RPPI - New dwellings for all types of dwellings (single- and multi-unit[1]) for urban areas[4]	RPPI New - Urban areas[4]	Sales and/or Stock based RPPI[2]	×		×	×	Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period
I.2.2.a. RPPI - New dwellings for single-unit dwellings[1] for the whole country	**RPPI New - single-unit dwellings[1] - Whole country**	**Sales and/or Stock based RPPI[2]**	×		×		**Whole country**	**At least quarterly**	**2005 Q1**	**Index - National reference period**
I.2.2.b. RPPI - New dwellings for single-unit dwellings[1] for the capital city[3]	RPPI New - single-unit dwellings[1] - Capital city[3]	Sales and/or Stock	×		×		Capital City	At least quarterly	2005 Q1	Index - National reference period

Name	Short name	Type					Coverage	Frequency	Start	Index
		based RPPI[2]								
I.2.2.c. RPPI - New dwellings for single-unit dwellings[1] for urban areas[4]	RPPI New - single-unit dwellings[1] - Urban areas[4]	Sales and/or Stock based RPPI[2]	×		×		Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period
I.2.3.a. RPPI - New dwellings for multi-unit dwellings[1] for the whole country	**RPPI New - Multi-unit dwellings[1] - Whole country**	**Sales and/or Stock based RPPI[2]**	×			×	**Whole country**	**At least quarterly**	**2005 Q1**	**Index - National reference period**
I.2.3.b. RPPI - New dwellings for multi-unit dwellings[1] for the capital city[3]	RPPI New - Multi-unit dwellings[1] - Capital city[3]	Sales and/or Stock based RPPI[2]	×			×	Capital City	At least quarterly	2005 Q1	Index - National reference period
I.2.3.c. RPPI - New dwellings for multi-unit dwellings[1] for the urban areas[4]	RPPI New - Multi-unit dwellings[1] - Urban areas[4]	Sales and/or Stock based RPPI[2]	×			×	Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period
I.3.1.a.RPPI - Existing dwellings for all types of dwellings (single- and multi-unit[1]) for the whole country	**RPPI Existing - Whole country**	**Sales and/or Stock based RPPI[2]**		×	×	×	**Whole country**	**At least quarterly**	**2005 Q1**	**Index - National reference period**
I.3.1.b. RPPI - Existing dwellings for all types of dwellings (single- and multi-unit[1]) for the capital city[3]	RPPI Existing - Capital city[3]	Sales and/or Stock based RPPI[2]		×	×	×	Capital City	At least quarterly	2005 Q1	Index - National reference period
I.3.1.c. RPPI - Existing dwellings for all types of dwellings (single- and multi-unit[1]) for urban areas[4]	RPPI Existing - Urban areas[4]	Sales and/or Stock based RPPI[2]		×	×	×	Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period
I.3.2.a.RPPI - Existing dwellings for single-unit dwellings[1] for the whole country	**RPPI Existing - single-unit**	**Sales and/or Stock**		×	×		**Whole country**	**At least quarterly**	**2005 Q1**	**Index - National reference period**

		dwellings[1] - Whole country	based RPPI[2]								
	I.3.2.b.RPPI - Existing dwellings for single-unit dwellings[1] for the capital city[3]	RPPI Existing - single-unit dwellings[1] - Capital city[3]	Sales and/or Stock based RPPI[2]		×	×		Capital City	At least quarterly	2005 Q1	Index - National reference period
	I.3.2.c.RPPI - existing dwellings for single-unit dwellings[1] for urban areas[4]	RPPI Existing - single-unit dwellings[1] - Urban areas[4]	Sales and/or Stock based RPPI[2]		×	×		Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period
	I.3.3.a.RPPI - existing dwellings for multi-unit dwellings[1] for the whole country	**RPPI Existing - Multi-unit dwellings[1] - Whole country**	**Sales and/or Stock based RPPI[2]**		×		×	**Whole country**	**At least quarterly**	**2005 Q1**	**Index - National reference period**
	I.3.3.b.RPPI - Existing dwellings for multi-unit dwellings[1] for the capital city[3]	RPPI Existing - Multi-unit dwellings[1] - Capital city[3]	Sales and/or Stock based RPPI[2]		×		×	Capital City	At least quarterly	2005 Q1	Index - National reference period
	I.3.3.c.RPPI - Existing dwellings for multi-unit dwellings[1] for urban areas[4]	RPPI Existing - Multi-unit dwellings[1] - Urban areas[4]	Sales and/or Stock based RPPI[2]		×		×	Urban areas[3]	At least quarterly	2005 Q1	Index - National reference period