INSTITUTO NACIONAL DE ESTADÍSTICA

Housing Price Index. Base 2015

Methodology

Sub-directorate General for Price and Household Budget Statistics

Madrid, June 2017

Index

1. Introduction

The Housing Price Index (HPI) was implemented in 2008 with the aim of completing the information provided by the existing statistics on the real estate market in Spain, with an approach based on the quarterly evolution of contract prices of purchase and sale of real estate.

Until the publication of the HPI, the main statistic produced by the public administrations was that which is now known as Housing Assessed Value, which is managed by the General Direction of Architecture, Housing and Land, of the Ministry of Public Works. The objective of this operation is to provide estimations on the average value per square metre of housing, on a quarterly basis. Its main source are the valuation companies, therefore, the published information refers to housing that has been appraised during the reference period, without needing to have changes produced in the type of housing that make up the sample, nor their quality, from one period to another, an aspect of particular emphasis for the HPI.

On the other hand, the information from the private sector, came from real estate sites and valuation companies, which also offered information on housing prices, specifically, on prices of the housing offered and appraised values. However, none of them was based in prices of transactions actually made.

This lack of information on purchase and sale prices, along with the need to process the changes in the sample of dwellings purchased in different periods, were the main reasons which prompted the National Statistics Institute (INE) to consider the realisation of a statistic that covered both aspects. In this way, the HPI uses prices of purchases and sales and employs a mixed adjustment method combining stratification with statistic methods of hedonic regression, which was an innovative and differentiating aspect with rest of existing official statistics.

Moreover, within the European Union, the interest of measuring the evolution in the housing prices dates back to de year 2002, when a Pilot Study Group concerned with home ownership was created, within the scope of the Harmonised Index of Consumer Prices (HICP). Spain has taken part in this group from its beginning. The work done within the group culminated with the publication of the HPI in Spain. In this sense, the HPI has been designed following the guidelines and recommendations of the European Statistical Office (Eurostat) which gives this indicator the possibility of establishing international comparisons.

Therefore, the implementation of the HPI met the information needs from the national and international scope, inasmuch as it is indispensable to have a harmonized indicator that makes it possible to establish international comparisons.

In order to do this, a working plan was designed with the purpose of implementing a Housing Price Index (HPI) bearing in mind the work carried out at the heart of the Eurostat Study Group.

Furthermore, in 2005 the INE set up an Inter-ministerial and Bank of Spain Work Group, with Presidency of the Government, Bank of Spain and Ministries of Economy and Tax, Justice and Housing representatives. This Working Group presented and analysed all of the elements that should comprise an indicator of the evolution of housing prices, with the basic characteristics required to enable the international comparison.

Eurostat began publishing the Harmonized Housing Price Index, as an experimental index, in December 2010. Subsequently, following the adoption of the Regulation (EU) No. Nº 93/2013, the official publication of the *Housing Price Index (HPI)* was initiated. The only differences with the previous HIP are that in the harmonized HPI the prices of new housing incorporate the VAT, as established in the Regulation, and that from 2016 the weightings reference period is two natural years in the older HPI and one natural year in the harmonized HPI.

The existing methodology details the procedure followed in the years previous to the implementation of the HPI and describes the most relevant characteristics of the indicator, the objectives and areas of the investigation, the source of information and methodology of the calculation used in its implementation and dissemination. The methodological section, the most extensive section, resumes the production process of the index and explains in detail, the regression model employed in the price estimates.

2. Background

The creation of the Study Group on home ownership coordinated by Eurostat served to promote the development of a new operation intended to measure the evolution of housing prices in Spain. In fact, the work carried out within this Group may be considered to be the preliminary on the work project followed in order to obtain the HPI.

In the **first phase** of the project (between the years 2002 and 2004) this Group, comprising Germany, Spain, Finland, Poland and the United Kingdom, had the objective of studying the situation of the real estate market in each of the countries, to find the main sources of information available and proposing a statistical calculation procedure that would enable making comparisons in the evolution of new housing prices among the EU countries.

During this phase in Spain, a pilot test for the collection of new housing prices was run, which had the main objectives of verifying the functioning of a collection in the field, and of testing the questionnaire model, which in the end analysed the differences existing between this information collection system and the obtaining thereof using valuation companies.

With this objective, the field test was circumscribed to the provinces of Madrid and Segovia, where new housing developments were visited in order to collect the sale prices and features of all types of housing within the development.

Likewise, preliminary studies were carried out, using appraisal values, using the average appraisal values made by the most important valuation companies in Spain in eight provinces, distinguishing by the size of the dwelling.

The conclusions obtained from these first tests served to begin to work on the design of the future index. Using these, the basic ideas were conceived for developing the index methodology, such as the ideal calculation periodicity, the cost/efficiency relationship of each one of the methods, and other relevant aspects, such as the processing of changes in quality of the dwellings.

Subsequently, in the **second phase** of the Study Group (which was developed in 2006 and 2007), another seven countries joined it: Cyprus, Slovenia, France, Greece, Italy, the Netherlands and the Czech Republic. The objective of this phase was to obtain data regarding the evolution of new and second-hand housing prices, in each one of these countries. The results were to be based on comparable calculation methodologies, which enabled constructing an owned housing price index with the highest degree of harmonisation. This is the main reason why the project for implementing the HPI had as a reference the general guidelines set out by Eurostat for the harmonised indicator.

Moreover, each country had to face the commitment of calculating an index of costs associated with housing acquisition, studying and selecting for it those items representative of that expense.

During this period, the INE analysed all of the available sources for obtaining the information necessary for calculating the HPI, on the one hand considering the availability, punctuality and content of the information, and on the other hand considering the demands of Eurostat regarding it, which required that the prices used to calculated the indices be the prices actually paid of the purchases and sales carried out during the reference period of the index.

Each one of the available sources of information regarding the housing market offers a different perspective on the purchase and sale process. In addition to the two sources previously mentioned (appraisals and real estate developments), there are other agents that provide information regarding housing: deeds, mortgage loans, Land registers and real estate agents.

After a detailed study of said sources, and having weighed the usefulness and convenience of each source for the established objectives, it has been decided that deeds values provide the best information for monitoring of owned housing prices, under the parameters of comparability and standardisation required by Eurostat.

There are numerous reasons why the remaining available sources have been dismissed for use in the project. In the case of appraisal values, the value of the dwelling estimated by the companies needs not be the same as the transaction price; likewise, not all of the dwellings appraised are subject to purchase and sale, nor are the dwellings whose payment has been made in cash–without a mortgage loan– considered.

In turn, the mortgage loans have been dismissed because the values thereof might not coincide with the transaction price, and this difference between the actual price and the value of the loan varies in each case.

The data from the Land Registry has also been dismissed, because the cadastral value does not provide an adequate measurement of the price of the dwelling, for the objectives set out by the HPI.

Lastly, despite the fact that the collection of prices using a survey targeting real estate developers could offer useful and complete information towards completing the project, this would only cover new housing, which would necessitate the use of other sources in order to obtain the prices of second-hand dwellings. The complexity and cost of said operation made it unfeasible.

Of all of these sources, that which was eventually selected for the calculation of the HPI was the notary register, which contains, among other data, the official prices for all of the purchases and sales occurring in Spanish territory, and correspond to the value of the property deed of the dwelling.

Moreover, the use of the administrative registers enables having available the information for the total population comprising the study or research scope, which favours the precision of the results and reduces the costs, if compared with other statistics that use sampling techniques for the field collection of the data.

The section on the source of the information details the most important features of the data base of the notaries. In 2007, this database experienced an in-depth restructuring and expansion of content, in order to meet the information demand of the Tax Agency. The HPI began to be published in October 2008, once the new database was consolidated. The indices published up to 2016 take 2007 as base year, the first year available with the new restructuring.

In parallel to the study of the sources, during this second phase, the ideal calculation method was decided for the housing price index, which would solve the problem of the changes in quality and in the composition of the sample of dwellings used each quarter to calculate the index. This method requires the use of regression models, as will be seen in detail in the section dedicated to the calculation methodology.

On the other hand, within Eurostat the working plan of the Study Group was continued, in which all the EU countries ended up participating, plus Norway and Iceland. The objectives were broadened and the quality of the indices improved, both with regard to their coverage as to the calculation method, until achieving the adequate level of comparability and harmonisation between countries in order to begin its dissemination. Thus, in 2011, Eurostat began to publish, on an experimental basis, a housing price index for the total of dwellings initially, and for new and second-hand dwellings afterwards. Independent data by country and aggregate data were provided, both for the Monetary Union and the European Union.

Two years later, following the adoption of the European Commission Regulation Nº 93/2013 of 23 February 2013, began the official publication of the *Housing Price Index (HPI)*, a harmonised index which differs slightly from the HPI published by the INE. The main difference concerning to the prices of new dwellings is that, as required by the European regulation, the harmonized index includes VAT. In addition, since 2016, the reference weighting period differs in both indicators. While the previous HPI uses the purchases and sales made during the two calendar years prior to the index reference year, the harmonized HPI uses those made during the last calendar year, as required by the Regulation.

3. Objectives

The HPI is a quarterly indicator whose main objective is to measure the **evolution of the level of the purchase and sale prices of new and second-hand free price housing** over time. This is therefore an indicator conceived solely for establishing comparisons over time.

The measurement of price levels does not fall within its scope. Therefore, it will not be possible to formulate spatial comparisons of price levels, but it will be possible to do so of their evolution.

In accordance with the criteria on the coverage of the Harmonised Index of Consumer Prices (HICP), protected dwellings are excluded from the calculation of the HPI, because they are not accessible to all possible buyers. These are dwellings whose typology, dimensions and prices are regulated by the Administration, as a condition for being able to apply for certain economic and tax advantages on the part of the buyers, which in turn must meet some established conditions relating to the ownership of property, family, income, etc.

Similarly, within the scope of production of European Union harmonised statistics, this index aims to serve as a comparative element between EU Member States, insofar as housing price evolution is concerned. In this sense, it has been conceived using the same concepts and methodology as those used in the production of the HCPIs of the EU. Nevertheless, as mentioned in the previous section, there are some conceptual differences between the European indicator and the HPI; being the most important the incorporation of VAT to the prices of new dwellings in the HPI.

4. Research scope

Statistical units

Bearing in mind that the objective of the HPI is to measure the evolution of the level of purchase and sale prices of free price housing in Spain, the unit of analysis is the free price housing. The basic statistical unit, on which the information is collected, is the housing. However, during the calculation process the housing typologies are established, for which the basic indices are calculated. Housing typology is therefore the *statistical derived unit*, which comprises housings with similar physical and localisation characteristics.

The *reporting unit* is the individual who provides the data of the property transfer to the notary. All the notarial information is centralised and the Notarial Certification Agency (ANCERT) is responsible for providing the data to the different users, among which is the INE.

The *concept* to which the index refers, is the housing price.

Population scope

The HPI population or reference stratum includes the entire population (individuals), both resident and non-resident in Spain, who have acquired a dwelling during the reference period. Those purchases made by legal persons or financial institutions are not included in the population scope of the HPI.

Geographical scope

The geographical scope of the research is comprised of the entire national territory.

Time scope

The HPI is published on a quarterly basis.

5. Source of information

The information used for calculating the HPI is taken from the General Council of Notaries, with whom the INE has signed a partnership agreement aimed at enabling the use of data from notaries for statistical purposes. Pursuant to this, the General Council of Notaries, via the Notarial Certification Agency (ANCERT), provides the data making up the main source of information for this indicator.

This source of information is the one best suited to the HPI objectives. The main characteristics that make it the ideal source are the following:

Availability

The agreement signed by the General Council of Notaries and the INE allows to dispose of the information necessary for the calculation of the HPI with the appropriate periodicity and timeliness for this statistical operation.

Periodicity

ANCERT provides on a monthly basis, in electronic format, information on real estate property transfers made in Spain during the previous four months. With each submission, the previously provided files are updated and new observations and modifications of the previous ones are incorporated.

Coverage

For the quarterly calculation of the HPI, the data files received approximately one and a half month after the end of the index reference quarter are used. In this way, more than 90% of the total transactions carried out in the reference quarter of the index are included, which to a great extent meets the needs of the indicator. This includes information regarding all the purchase and sales carried out in the national territory throughout the quarter, of both new and second-hand dwellings, and regardless of the form of payment, in cash or with credit.

Timeliness

The prices that should be part of the calculation of the index are those that refer to the purchase and sales or dwellings actually made during the reference quarter. Therefore, neither offer prices nor appraisal values are included. In this manner the criterion of timeliness is met, which requires that transactions made should be included in the index calculation in the same quarter in which they take place.

Content

The database of notaries, in addition to the purchase and sale price of the dwelling, contains additional very valuable information that allows, on the one hand, the adaptation to the coverage of the index in terms of kind of transfer, type of property and buyer, but also the implementation of a good quality adjustment by having detailed information on the features of the dwelling, among them, its size and location, which are more relevant aspects in the determination of the price.

Moreover, it counts with the reference of the land registry of the property, from which cadastral information can be accessed and used; for the moment it is used during the cleansing phase of the survey.

6. Variables used

Variables included in the data files which ANCERT submits periodically are showed below, grouped according to different aspects of interest:

- **Temporary variable:**
  - Date of authorisation. Indicates the day in which the transfer is made.
- Operation code:

The transactions corresponding to the following legal acts are received:

- Purchase and sale of real estate property.
- Allocation of cooperative housing to its members.
- Allocation to co-proprietor in the real estate development community.

- **Variables relating to the price of the dwelling:**
  - Price of the operation.
  - Value of the object.

In the acts with more than one object, that is, when the same legal act involves the transfer of multiple objects (for example, a dwelling and a parking space), and the *price of the operation* includes all of them, a value is assigned to each one of them in the *value of the object* field.

- **Dwelling location variables:**
  - Autonomous Community.
  - Province.
  - Municipality.
  - Postal code.
  - Type, name and number of the road.
  - Duplicate, block, stairwell, floor and door.
- **Variables relating to features of the dwelling:**
  - Type of real estate property. This indicates the type of urban property: dwelling, parking space, storage room...
  - Reference in the land registry of the property, or motive for its absence or reason for not being able to obtain it, where appropriate.
  - Variable indicating whether the dwelling is free-price or with official protection.
  - Type of dwelling, distinguishing between flat and single-family dwelling.
  - Variables indicating whether the parking space and storage room are included in the price.

- Variable indicating whether the dwelling is new or second-hand.
- A dwelling is deemed to be new when it is the first transfer on the deed of sale, which is normally carried out by the developer or construction company on behalf of the first buyer; for the remainder of transfers, in other words, when there is more than one transfer on the public deed, the dwelling is deemed to be second-hand.
- Built-up area in square metres (m2).

– **Variables referring to the buyer:**

- Type of person. This indicates if the buyer is a natural or legal person.
- Country, province and municipality of residence of the buyer.

All variables previously mentioned intervene, directly or indirectly, in the production process of the HPI, with exception of those related to the place of residence of the buyer, since the HPI includes all dwellings acquired by individuals within the Spanish territory, regardless of their nationality or place of residence.

As pointed out above, the location of the dwelling, together with the size, are the most relevant elements in the explanation of the price. The database includes the exact address of the property; however, in order to use this information in the regression model which is employed in the estimation of prices, it is necessary to group together the provinces, municipalities and postcodes, and thus have a number of categories not too high to avoid the over-parametrisation of the model. To achieve this, other sources of information have been used, which have allowed the creation of new variables, classification of the different geographic levels, according to one or several variables that have relation with the price of the dwelling. These variables are the following:

– **Cluster of provinces**. Cluster of the 52 provinces in 6 groups through the application of a cluster analysis, using the average annual appraisal value of the dwelling by province, published by the Ministry of Public Works.

– **Size of the municipality**. Classification of the municipalities, distinguishing between provincial capitals and large non-capital municipalities (with more than 50,000 inhabitants), medium-sized municipalities (with 10,000 to 50,000 inhabitants) and small municipalities (with fewer than 10,000 inhabitants), using the latest population data available from the Continuous Municipal Register (INE).

– **Tourism-oriented municipality**. A municipality is considered to be tourism-oriented if it concentrates a high number of overnight stays in tourist accommodation sites and/or the proportion of secondary residences against principal residences is high. These are classified in any of the first three categories:

- Sun and beach tourism
- Rural, inland or natural tourism
- Cultural, urban or business tourism
- Rest (non-tourist)

For this, information on the number of annual overnight stays in each type of tourist establishment is used, provided by the INE surveys for Hotel Occupancy, Holiday Dwellings Occupancy and Rural Tourism Accommodation Occupancy, as well as data on principal and secondary residences from the last housing census, the population of the municipalities and its location (coast or inland).

- **Type of environment**. Classification of postcodes in 14 categories, based on information from the last housing census (2011) and from the corresponding price per square metre by postcode, in an earlier annual period.

Every year, the variables classifying the provinces, municipalities and postcodes are updated with the latest available information from the sources used in their compilation, keeping or not the number of categories and/or their content. For example, the variable *tourism-oriented municipality* had only two categories, tourism-oriented or not tourism-oriented, considering only the criterion of high number of overnight stays. The initial years, the provinces were grouped according to the average mortgage cost; and the limit to distinguish medium-sized municipalities from the large ones was 100,000 instead of 50,000 inhabitants.

The use of chain indices enables to introduce modifications annually, without there being an appreciative effect in the variation rates of the indices.

7. Data processing

The files provided by the General Notary Council must be adapted to the technical requirements needed for the calculation of the HPI. For this purpose, a process has been designed which guarantees the internal consistency of the data and excludes extreme values.

The process consists of the following phases:

Initial phase

During the initial phase of the process, the three monthly files which comprise the quarter are combined, all the variables received are formatted, and a first filter is made to adapt the records to the different areas of the indicator; in order to do this, the purchase and sales realized by natural persons which refer exclusively to free price dwellings are selected. The resulting file should include only dwellings, so that multiple transfers–when in a same act different properties such as parking buildings and storage rooms are jointly transferred–must pass through a specific process which break down the values of each one of them.

Also during this first phase the significantly high and low values are eliminated, both in the area as well a in the price of the dwelling. Afterwards, beginning with the regression model, the atypical observations will be identified in an objective manner and taking into account all the characteristics as a whole of the dwellings collected in the model.

The dwellings excluded due to having an atypical value in the surface variable will be analysed in a later phase. If the issue is an error in values and their correction is available, they will be incorporated in the study again.

Filtering and Imputation Phase

The second phase of the process focuses on the imputation and cleansing of values. It consists in the detection of inconsistent values of the geographical variables (for example, a postcode which does not belong to the saved municipality). These errors are detected automatically and are resolved recurring to external sources.

Even though the notarial data bases are usually complete, in some cases it is necessary to allocate the value of variables such as surface or type of environment. In the case that it should be the surface area, the cadastral information for dwellings will be used to compare and correct the value recorded, in case of an error.

In as far as the type of environment, as it has already been described in the previous section, it would be a classification for the postcodes; however, not all of them are classified. Therefore an allocation process has been designed based on the observation of the average quarterly prices: to the observation whose postcode is not classified it shall allocate the type of environment of that postcode, which being within the same municipality, has the average price per square metre more similar to the price of the square metre of the observation in the corresponding quarter.

Expansion Phase

Finally, even though some derived variables are obtained during the process, as it occurs, for example, with the price per square meter (price by area), which is necessary in the allocation phase, the majority of them are incorporated to the data file at the end. As is the case of many explanatory variables of the regressive model.

As it has been mentioned before, the location of the dwelling, together with its size, are the factors which most determine the explanation of its price. The data base counts with the exact address of the dwelling; however, in order to use it in the regression model, it is necessary to abbreviate this information into a few variables with a reduced number of values each, thus avoiding the over-parametrisation of the model. For this reason, the provinces, the municipalities and the postcodes have been grouped taking into account information of other sources about some variable closely related with the price of the dwelling; in the previous section there is an explanation on how these variables were created. Also with the same objective, 10 surface intervals have been created and the values of the floor variable–which collects the hight of the flat within the building– have been grouped into six categories.

The file which results from this process is completely adapted to the coverage of the HPI and prepared to–based on a regression model described the following section– obtain the estimated prices that will entry into the calculation formula of the index.

8. Calculation methodology

The calculation system of the HPI is based on the combination of two basic elements that reflect the characteristics of the real estate market, and which are essential to the calculation of the price indices: the prices of the dwellings, which represent the confluence of market supply and demand, and the weightings, or relative importance of each type of dwelling according to the value of the purchase.

The combination of these two elements in order to obtain the HPI is carried out using the **chain-linked Laspeyres index formula**, the same that is used in calculating the CPI/HICP.

In addition to the consideration of the two previously mentioned elements, another of the relevant aspects in any price index is the adjustment due to the change occurring in the quality of the goods whose prices are followed over time. When the prices observed correspond to dwellings, this aspect is of the maximum importance. In this case, the monitoring of the price of the same dwelling every quarter is impossible; in fact, the composition of the set of dwellings used for the calculation of the index is different each quarter, since it is comprised of those dwellings sold in said quarter. Therefore, if the prices are not adjusted due to the change in the composition of the sample or the quality of the dwellings, the estimation in the evolution thereof would not be representative of the real estate market.

A possible solution is the grouping in strata of dwellings with similar features. In this way, the average price in each stratum is more representative, given its homogeneity. Logically, in order to make more precise the evolution of the prices, it is convenient that the strata be reduced, since the more delimited is the typology of the dwelling, the more efficient will be the adjustment for change of quality and composition.

On the other hand, in order to obtain representative average prices for each stratum with the traditional estimators, it is necessary to have a minimum number of observations per stratum each quarter. This demand would require a decrease in the degree of detail in the stratum, reducing the number of characteristics that define it. As a result, those dwellings belonging to the same stratum would not be as homogeneous as would be hoped for. For this reason, the HPI uses a mixed method which combines stratified and hedonic techniques, allowing to estimate prices for each stratum, regardless of the number of dwellings belonging to it in the quarter. In this way, the number of typologies considered, and the degree of detail in the definition thereof, is greater, which significantly improves the adjustment.

Hedonic models are commonly used in the calculation of price indices to monitor changes in the quality of products that make up the indices. Those models intend to explain the value of a good in terms of each one of the attributes or characteristics that compose it. With this is possible to determine how this value will change by varying the quantity in which each one of these attributes is present, and consequently, to predict prices.

The following presents, with the greatest detail, the calculation process of the HPI, affecting the mentioned aspect, and in a particular way, affecting the regression model used.

8.1 Prices

As stated previously, the prices per metre squared are one of the basic elements in the calculation of this indicator. However, given the heterogeneity of the dwellings, it is necessary to apply, to these prices, a process that guarantees the comparability thereof; therefore, the prices that are eventually involved in the computation of the HPI are those that are obtained for each stratum or typology of dwelling after applying the estimation process.

To the received data, the process described in the previous section is applied, to obtain records related to dwellings and adapted to the coverage of the HPI, with complete and cleansed information. The regression model is applied on this final dwellings file to obtain the estimated coefficients of the model, that reflect the implicit prices of the characteristics of the dwellings. Through them, the prices for each typology of dwelling, which are those involved in the calculation of basic indices.

Even though the prices are estimated on a quarterly basis with the sample of purchase and sales of dwellings made in the quarter, the set of housing types is fixed for an entire year. Each combination of values of the 11 variables included in the regression model (which will be seen in the next section), present in any of the transactions carried out during the weightings reference period, constitute a housing typology. As of 2016, the weightings reference period is the previous two years. In 2017, approximately 49,000 different housing typologies have been constructed, and which have been determined by observing the physical characteristics and location of the dwellings subject of the transaction during the years 2015 and 2016.

The prices of each typology are estimated quarterly, using the information provided on purchases an sales of dwellings carried out during the quarter, regardless of the number of quarterly transactions for each typology. Here lies the advantage of the method employed, stratified with hedonic. The model allows to get an estimated price for all housing typologies, which is obtained multiplying the vector of characteristics that defines each housing typology by the vector of parameters, which changes on a quarterly basis and reflects the implicit prices of the characteristics.

With the prices estimated by the model, the basic indices for each typology are calculated, which together with their weighting are used for the calculation of the aggregate indices.

8.2 Regression model

The regression model applied for the calculation of the HPI is a semi-logarithmic model–commonly used in this field–where the dependent variable is the naperian logarithm of the price per square metre of the dwelling.

The explicative variables, which include the physical characteristics and location of the dwelling, are all categorical, that is, they take a finite number of values.

The following table includes the eleven main effects or explanatory variables of the model, with their correspondent values or categories. Except the floor variable, which began to be used in 2010, all of them has been part of the model since the beginning of the publication.

Explanatory variables of the regression model used in the HPI

Variable	*values*	*Categories or values*
New/Second-hand	2	A dwelling is considered new when it is related to the first transfer.
Type of housing	2	Flat or single-family dwelling
Garage	2	Yes or no
Storage room	2	Yes or no
Cooperative	2	Yes or no
Area	10	<40 m2; [40,60); [60,75); [75, 90); [90, 105); [105, 120); [120, 150); [150, 180); [180, 240); ≥240 m2
Floor	6	Basement, ground floor, first, second, rest of floors and top floors.
Classification of provinces	6	With information on the average appraisal value by province 6 groups of provinces are established.
Size of the municipality	4	Capitals, non-capital municipalities with more than 50,000 inhabitants, between 10,000 and 50,000, and less than 10,000 inhabitants.
Tourism-oriented municipality	4	Tourism-oriented municipalities are considered those that concentrate a high number of overnight stays in tourism accommodation sites and/or a high percentage of second residences. These are classified in any of the first three categories: • sun and beach tourism • rural, inland or natural tourism • cultural, urban or business tourism • rest (non tourism-oriented)
Type of environment	14	Grouping of postcodes in 14 types of environment based on census information and information on the average price per square metre in the previous year.

The first five variables are dichotomous and are obtained directly from the data file of the notaries; the last six have been created grouping the values of some of the received variables. For this, in some cases, it has been necessary to use information from other sources, as described in detail in the section on the data processing, where the last four variables have been defined.

Moreover, the model includes the most significant double interactions between these main effects. The criteria followed for selecting the interactions are the following three: they must be significant, their contribution to the explicative power of the model must be as high as possible, and the number of quarterly observations must be greater than 30 in each one of the combinations or pairs of possible values of the interaction. Each interaction adds restrictions to the model, specifically the observed average price and the estimated by the model should coincide at each crossover; hence the need to require a minimum number of quarterly observations in the third criterion.

The number of interactions of the model has remained at nine; however some of them might have change from one year to the next. As a result, the number of parameters of the model has also changed annually, between 120 and 150. In this way, with the estimated parameters every quarter, the prices of thousands of different typologies are estimated, which underlines again, the advantage of the method employed.

Both the main effects and the interactions can change annually, as the model will be subject to review each year. This review consist of the following:

- Updating the variables related to the location with the latest information available from the sources used in its elaboration. Thus, from one year to another the content

can vary (and even the number) of categories of variables: *cluster of provinces*, *size of the municipality*, *tourism-oriented municipality* and *type of environment*.

- Adding new possible explanatory variables from the notaries' databases, or created from complementary information, as happened in 2010 when the variable floor was incorporated, after analysing and grouping the values recorded in that field in the database.

- Reviewing the model interactions. For this, the fulfilment of established criteria with the data of the last four quarters is observed.

Initially, the regression model is applied on the final data, obtained after the processing of the information. Then, the atypical observations are selected and eliminated from the residues of the initial model. It is, therefore, an objective way which takes into account, both the price and the value of the 11 explanatory variables of the model. Finally, the final model, with which the prices involved in the index calculation are estimated, is a weighted model, in order to, on the one hand, correct the heteroskedasticity, and on the other, to assign a weight (less than one) to those observations where the value of any variable has been imputed. The model specification with all the formulation and technical information necessary for the estimation of prices is described in the annex II.

8.3 Weightings

The weighting structure enables establishing the importance or weight that each stratum or typology of dwelling has in comparison with the others, depending on the expenditure made in the purchase of each type of dwelling with regard to the total expenditure in the purchase of dwellings during the reference period. This is thus a flow variable–transactions carried out–and not a stock variable such as the number of owned dwellings existing in Spain.

The source of information used for obtaining the weightings is the same as that used to obtain the prices, as the notary data enables ascertaining the types of transactions carried out over time, both in number and in value.

Due to the dynamism of the real estate market, it is convenient to annually update the weighting structure in such a way that it represents the status of the market as reliably as possible. The chained Laspeyres formula which the HPI uses allows to update the weightings every year.

As in all the indices designed from a chained index scheme, any change introduced from one year to the next has some effect in the annual variation rates of the index. However, this obstacle is compensated by the permanent adaptation of the indicator to the changes occurred in the market. In this permanent adaptation, apart from looking for topicality, it should also guarantee a certain level of stability to the weighting structure. In this sense, the more years involved in the calculation of the weights, the less they will fluctuate and their effect on the annual price variations will be lower. Moreover, the variety of dwelling typologies will also be greater, which will improve the adjustment for change of quality and composition.

Initially and until 2013, the information of three-year transactions was used to obtain the weighting structure of the HPI. With the entry into force of the European regulation

for the harmonised housing price index, which established the use of a single year for the calculation of weights, in the same way as the HCPI, the HPI also began to use a single year for the calculation of the annual weighting structure. However, since 2016, two years of information are used in the calculation of weights with the aim to maintain the topicality of the indicator and provide it with a consistent structure. Thus, the current weights in 2017 have been obtained with data on the transactions carried out during the years 2015 and 2016.

The HPI uses hybrid weights; the weights are said to be hybrid when the quantities in one period are valued at the prices of another period. The weight calculation formula of a housing typology or stratus *e*, in the year *a*, is the following:

$$W_e^a = \frac{Q_e^{(a-1,a-2)} \times \hat{P}_e^{\,4,\,a-1}}{\sum_{\forall e} Q_e^{(a-1,a-2)} \times \hat{P}_e^{\,4,\,a-1}} \qquad a \geq 2016$$

where, both prices and quantities refer to the same unit, the housing square metre. Thus:

$Q_e^{(a-1,a-2)}$ represents the average annual number of square metres of dwellings belonging to stratum *e*, sold during the reference period of weightings *(a-1, a-2)*, and

$\hat{P}_e^{\,4,a-1}$ is the estimated price of the square metre estimated by the regression model for the stratum *e* in the fourth quarter of the previous year.

In this way, the annual weighting of each housing typology represents the expenditure incurred in the purchase of this type of housing compared to the total number of housings, carried out in the previous two years, and valued at the prices of the fourth quarter of the previous year.

The fact of using estimated prices rather than collected prices is due to the fact that, during the fourth quarter of the year (*a-1*), it is possible that dwellings from all of the typologies have not been sold, and therefore, information is not available for observed prices for all of the strata. Furthermore, according to the chained Laspeyres formula, the prices involved in the calculation of the basic indices should be the same used in the calculation of the weights; in both cases, the HPI uses the prices estimated by the model.

The weighting of any aggregate *A*, whether it be functional or geographical, is obtained as the sum of the weightings of the strata comprised by said aggregate:

$$W_A^a = \sum_{e \in A} W_e^a$$

8.4 Calculation of indices

The general formula used in the calculation of the HPI is a chained Laspeyres index, analogous to that used in the CPI/HCPI. In the case of the HPI, since it is a quarterly

indicator, the period used for chaining is the fourth quarter of each year, and not the month of December.

The use of chained indices enables the annual updating of the weightings, as well as the possibility of making methodological changes (such as the review of the regression model or the inclusion of new housing typologies), unlike what occurs with a fixed-base Laspeyres index, in which both the weightings and the methodology remain fixed throughout the period in which the base is in force.

In a chained index, three reference periods are defined:

- **Reference period of the index or base period.** Period for which the average of the indices is made to equal 100. It usually involves an annual period. In the HPI, since 2017, the base year is 2015, and all the indices published since then have as a reference this period.

- **Reference period of the weightings**. This is the period referred to in the data used in calculating weightings.

Each year, the HPI weightings are calculated with the latest available information on the number of purchase and sales in a previous period, these quantities are valued at the prices of the fourth quarter of the previous year. The period used for obtaining the quantities has changed, from the three years that were used at the beginning of the publication, it was changed to a single year in 2013 and, since 2016, the previous two years are used. Thus, in 2017, the weighting reference period is the one composed of the years 2015 y 2016.

- **Reference period of the prices**. This is the period with whose prices the current prices are compared, that is to say, the period chosen for the calculation of the basic indices. This is the fourth quarter of the year immediately preceding the current year.

The following expresses the calculation formula of the basic indices and the aggregate indices, as well as the general calculation system of the chained indices.

BASIC INDICES

A basic aggregated is the component with the lowest level of aggregation for which indices are obtained, and in whose calculation weightings are not involved; the indices of these aggregations are called basic indices. In the HPI, the basic aggregation is the stratum that includes a single typology of dwelling.

The basic index of the stratum is obtained as the quotient of the price estimated by the model for the dwellings belonging to said stratum in the current period, and the price estimated from the fourth quarter of the previous year:

$$_{(4,\,a-1)}I_e^{q,a} = \frac{\hat{P}_e^{q,a}}{\hat{P}_e^{4,\,a-1}} \times 100$$

where,

$\hat{P}_e^{q,a}$ the estimated price per square metre for those dwellings belonging to stratum *e*, in quarter *q* of year *a*, and

$\hat{P}_e^{4,a-1}$ the estimated price per square metre for those dwellings in the stratum e, and in the 4th quarter of the year *a-1*. The estimation of this price has been carried out with the same regression model[1] used in estimating the price of the numerator.

AGGREGATE INDICES REFERRING TO THE FOURTH QUARTER

The calculation of the index of aggregation A, whether of a functional or geographical type, is carried out using the basic indices of those strata belonging to said aggregation and its corresponding weightings, in accordance with the following expression:

$$_{(4,\,a-1)}I_A^{q,a} = \sum_{c \in A} W_e^a \times {}_{(4,\,a-1)}I_e^{q,a}$$

where,

W_e^a the weighting of stratum *e*, as so much per one, valid during year *a*[2], and

$_{(4,\,a-1)}I_e^{q,a}$ the basic index of stratum *e*, in quarter *q* of year *a*.

The previous formula can be expressed, in an equivalent manner, as the quotient of average weighted prices by determined quantities, the same in the numerator and denominator.

$$_{(4,\,a-1)}I_A^{q,a} = \frac{\sum_{e \in A} Q_e \times \hat{P}_e^{q,a}}{\sum_{e \in A} Q_e \times \hat{P}_e^{4,\,a-1}} \times 100$$

This is the usual expression of the Laspeyres indices, where the current prices (numerator) are compared with those of the fourth quarter of the previous year (denominator), keeping the quantities constant.

Using the indices referring to the fourth quarter, the quarterly repercussions and those accumulated over the year are calculated.

INDICES IN BASE 2015

The indices in base 2015 are those that are published and obtained linking the indices referring to the fourth quarter of the previous year, according to the following expression:

[1] The regression model is reviewed each year, in such a way that the prices for the fourth quarter of each year *a* must be estimated in two different ways. On the one hand; the model in force in year *a* will be used to calculate the numerator of the basic indices for the fourth quarter of year *a*. On the other hand, the revised model in force in the subsequent year *a+1,* the denominators of the basic indices of the four quarters of year *a+1* will be calculated.

[2] Depending on the year, the weightings are obtained with the information on the purchases and sales carried out in the previous three years (up to 2012), in the previous year (from 2013 to 2015) or in the previous two years (from 2016).

$$ {}_{15}I_A^{q,\,a} = {}_{15}I_A^{4,\,(a-1)} \times \left(\frac{{}_{4,(a-1)}I_A^{q,\,a}}{100} \right) = \qquad a \geq 2016 $$

$$ = {}_{15}I_A^{4,\,15} \times \left(\frac{{}_{(4,15)}I_A^{4,\,16}}{100} \right) \times \ldots \times \left(\frac{{}_{(4,\,a-2)}I_A^{(4,\,a-1)}}{100} \right) \times \left(\frac{{}_{(4,\,a-1)}I_A^{q,a}}{100} \right) $$

Using the indices in base 2015, the quarterly variation rates are obtained, the accumulated (or year-to-date) variation rate and the annual variation rate. The two first rates can be also calculated from the indices referred to the 4th quarter.

CHAINED SERIES

In CPI base 2015, the only change has been the reference period of the indices or the base period, which went from the year 2007 to the year 2015. As a result, in base 2007, a re-scale coefficient has been calculated, which has converted the indices published in base 2007, from the first quarter of 2007 to the last quarter of 2016, into indices in base 2015.

This coefficient is that which makes the simple arithmetic average of indices published in the year 2015, in base 2007, equal to 100.

$$ \frac{1}{4}\sum_{q=1}^{4} {}_{07}I^{\,q07} \times C_{re-escala} = 100 \Leftrightarrow C_{re-escala} = \frac{100}{\frac{1}{4}\sum_{q=1}^{4} {}_{07}I^{\,q07}} $$

8.5 Calculation of variation rates

QUARTERLY VARIATION RATE

The quarterly variation rate of an index is calculated as the quotient between the index for current quarter and the index for previous quarter, both in base 2015, according to the following formula:

$$ \Delta^{qa/(q-1)a} = \left(\frac{{}_{15}I^{qa}}{{}_{15}I^{(q-1)a}} - 1 \right) \times 100 $$

where:

$\Delta^{qa/(q-1)a}$ the quarterly variation rate of prices in quarter *q* for year *a*, in percent,

${}_{15}I^{qa}$ the index of quarter *q* in year *a*, in base 2015, and

${}_{15}I^{(q-1)a}$ the index of quarter *q-1* of year *a*, in base 2015.

ACCUMULATED VARIATION RATE

The accumulated, or year-to-date, variation rate is calculated as the quotient between the index for the current quarter and the index from the fourth quarter of the previous year, both in base 2015:

$$\Delta^{qa/4(a-1)} = \left(\frac{{}_{15}I^{qa}}{{}_{15}I^{4(a-1)}} - 1 \right) \times 100$$

where:

$\Delta^{qa/4(a-1)}$ the accumulated variation rate of prices in quarter *q* of year *a*, in percent

${}_{15}I^{qa}$ the index in base 2015, in quarter *q* of year *a*, and

${}_{15}I^{4(a-1)}$ the index in base 2015, in the fourth quarter of year *a-1*.

ANNUAL VARIATION RATE

The annual variation rate is calculated as the quotient between the indices published for the current quarter, and for the same quarter the previous year, both in base 2015:

$$\Delta^{qa/q(a-1)} = \left(\frac{{}_{15}I^{qa}}{{}_{15}I^{q(a-1)}} - 1 \right) \times 100$$

where:

$\Delta^{qa/q(a-1)}$ the annual variation rate of prices in quarter *q* of year *a*, in percent

${}_{15}I^{qa}$ the index in base 2015, in quarter *q* of year *a*, and

${}_{15}I^{q(a-1)}$ the index in base 2015, in quarter *q* of year *a-1.*

8.6 Calculation of repercussions

QUARTERLY REPERCUSSIONS

The repercussion of the quarterly variation of a stratum or group of strata of dwellings in the general index, is defined as the part of the quarterly variation of the general index that corresponds to said stratum or group of strata. Therefore, the sum of the quarterly repercussions of all of the strata of dwellings included in the HPI, is equal to the quarterly variation of the general index.

In other words, the repercussion that the quarterly variation of prices of a stratum or group of strata has on the quarterly variation of the general index, is the variation that it would have experienced if all of the prices of the remaining strata had not varied during that quarter.

The formula for the quarterly repercussion of a given stratum (or group of strata), in quarter *q* of year *a*, is as follows:

$$R_e^{qa/(q-1)a} = \frac{{}_{4(a-1)}I_e^{qa} - {}_{4(a-1)}I_e^{(q-1)a}}{{}_{4(a-1)}I_G^{(q-1)a}} \times W_e^a \times 100$$

where:

${}_{4(a-1)}I_e^{qa}$ is the index referring to the 4th quarter of the year *a-1* for stratum *e*, in quarter *q* of year *a*,

${}_{4(a-1)}I_e^{(q-1)a}$ is the index referring to the fourth quarter of the year *a-1* for stratum *e*, in the quarter *q-1* of year *a*,

${}_{4(a-1)}I_G^{(q-1)a}$ is the general index referring to the fourth quarter of the year *a-1*, in quarter *q-1* of year *a*, and

W_e^a is the weighting in force in year *a* for stratum *e*, as so much per one.

As may be observed, the repercussions are calculated using the indices referring to the fourth quarter of the previous year (non-published indices). An alternative expression of the above formula is as follows:

$$R_e^{qa/(q-1)a} = \frac{{}_{4(a-1)}I_e^{qa} - {}_{4(a-1)}I_e^{(q-1)a}}{{}_{4(a-1)}I_G^{(q-1)a}} \times W_e^a \times 100 =$$

$$= \frac{{}_{4(a-1)}I_e^{qa} - {}_{4(a-1)}I_e^{(q-1)a}}{{}_{4(a-1)}I_G^{(q-1)a}} \times \frac{{}_{4(a-1)}I_e^{(q-1)a}}{{}_{4(a-1)}I_e^{(q-1)a}} \times W_e^a \times 100 =$$

$$= \Delta_e^{qa/(q-1)a} \times \frac{{}_{4(a-1)}I_e^{(q-1)a}}{{}_{4(a-1)}I_G^{(q-1)a}} \times W_e^a$$

Therefore, the quarterly repercussion of a given stratum e, is the product of its quarterly variation rate of prices in percent, $\Delta_e^{qa/(q-1)a}$, by its weighting as so much per one, W_e^a, and by the quotient between the index for the stratum and the general index, both for the previous quarter, $\frac{{}_{4(a-1)}I_e^{(q-1)a}}{{}_{4(a-1)}I_G^{(q-1)a}}$.

As was mentioned previously, the sum of the quarterly repercussions of all of the strata comprising the set of dwelling typologies of the HPI is equal to the quarterly variation of the general index, as demonstrated below.

$$\sum_e R_e^{qa/(q-1)a} = \sum_e \frac{{}_{4(a-1)}I_e^{qa} - {}_{4(a-1)}I_e^{(q-1)a}}{{}_{4(a-1)}I_G^{(q-1)a}} \times W_e^a \times 100 =$$

$$= \left(\frac{\sum_e {}_{4(a-1)}I_e^{qa} \times W_e^a - \sum_e {}_{4(a-1)}I_e^{(q-1)a} \times W_e^a}{{}_{4(a-1)}I_G^{(q-1)a}} \right) \times 100 =$$

$$= \frac{{}_{4(a-1)}I_G^{qa} - {}_{4(a-1)}I_G^{(q-1)a}}{{}_{4(a-1)}I_G^{(q-1)a}} \times 100 = \Delta_G^{qa/(q-1)a}$$

ACCUMULATED REPERCUSSIONS

The accumulated repercussion, or for the year-to-date, of a stratum or group of strata in the general index, represents the accumulated variation that the general index would experience if the remaining strata experienced no variation at all in prices for the year-to-date; in other words, it is the part of the accumulated variation due to said stratum or group of strata.

The formula for the accumulated or year-to-date repercussion of a given stratum *e* (or of a given aggregate) in quarter *q* of year *a*, is as follows:

$$R_e^{qa/4(a-1)} = \frac{{}_{4(a-1)}I_e^{qa} - {}_{4(a-1)}I_e^{q(a-1)}}{{}_{4(a-1)}I_G^{q(a-1)}} \times W_e^a \times 100 =$$

$$= \frac{{}_{4(a-1)}I_e^{qa} - 100}{100} \times W_e^a \times 100 = \Delta_e^{qa/4(a-1)} \times W_e^a$$

where:

$\Delta_e^{qa/4(a-1)}$ is the accumulated variation rate of stratum *e*, in quarter *q* of year *a*, in percent, and

W_e^a is the weighting of stratum *e* in force in year *a*, as so much per one.

Therefore, the year-to-date repercussion is the product of the accumulated variation rate (as so much per hundred) and the weighting (as so much per one). Thus, the higher the repercussion of a stratum or group of strata on the accumulated variation of the general index, the higher its accumulated variation and weighting.

In this case, the sum of the accumulated repercussions of all of the strata is equal to the year-to-date variation of the general index:

$$\sum_i R_e^{qa/4(a-1)} = \sum_e \left({}_{4(a-1)}I_e^{qa} - 100 \right) \times W_e^a =$$

$$= \sum_e {}_{4(a-1)}I_e^{qa} \times W_e^a - 100 \sum_e W_e^a = {}_{4(a-1)}I_G^{qa} - 100 =$$

$$= \frac{{}_{4(a-1)}I_G^{qa} - 100}{100} \times 100 = \Delta_G^{qa/4(a-1)}$$

9. Dissemination

El HPI is published quarterly approximately 70 days after the end of the index reference period, according to the calendar of the publications of the INE. In the fourth quarter of each year, the INE disseminates the calendar with the exact dates of publication of the statistics for the next year.

Each quarter, the INE publishes the press release of the HPI where the most important price variations are commented and the main results are shown.

INEbase is the system used by the INE to store and disseminate all statistical information on the website, in electronic format. The INE web page offers access to the on-line data base of the HPI, which provides information with respect to price indices, variation rates and weightings.

Data is published for the nation as a whole, the 17 Autonomous Communities and the Autonomous Cities of Ceuta and Melilla, which enables establishing comparisons between the evolution of prices in the different regions.

Regarding the functional breakdown,information is provided regarding new and second-hand housing, on a national level. Due to the demand of more detailed information on behalf of the users, since 2010, the breakdown of new and second-hand housing is also available my Autonomous Community.

The following data will be published quarterly:

- 2015 base indices (until 2016, in base 2007);
- quarterly variation rates;
- year-to-date variation rates;
- annual variation rates;
- quarterly repercussions (for new housing and second-hand housing);
- year-to-date repercussions (for new housing and second-hand housing).

Besides the quarterly results, each year the annual averages of the indices and the variations of the annual averages for each published series are given. In the same way, the annual weighting structure is made available to the users.

The HPI data is final data from the first time it is published and therefore it is not subject to revision.

The standardized methodological report, which is accessible from the web page of the INE, contains the metadata information of the survey, which help to understand and interpret the results better.

Annex I. Glossary of terms

- **ANCERT**. The Notarial Certification Agency is the technological Association of the Spanish Notarial created by the General Council of Notaries with the object of modernizing and placing all of the Notaries of Spain in the technological vanguard as well as the different bodies of the notarial collective.

- **Cell**. Combination of the possible values of the variables or characteristics (main effects of the regression model) defining a specific housing type.

- **General Council of Notaries**. Institution coordinating the Professional Associations of Notaries throughout Spain. It manages the database relating to real estate property operations (computerised index of notaries), which is used for calculating the HPI.

- **Housing cooperative**. This is the group of persons who, meeting the general requirements of the cooperative (establishing statutes, registering in the Cooperative Register, composition of the bodies by which it works, accounting, etc.), meet to participate in a common project, carrying out however many activities are necessary (search for plots, search for a financial institution to finance construction, hire the architect, write up incorporation contracts, building contract, dwelling allocation contracts, etc.) to achieve accommodation and/or complementary locales and installations, for themselves or for those persons who live with them.

- **Main effect**. Regression model explanatory variable.

- **Mortgage**. Entitlement contracted by the mortgage lender as compared with the borrower, in the case of non-payment of obligations by the latter, and which is exercised on the property appearing as a guarantee or collateral. In the case of a mortgage loan for a dwelling, the property mortgaged Is usually the dwelling purchased.

- **Interaction**. Regression model explanatory variable, obtained as a combination of other explanatory variables (main effects) of the model.

- **Hedonic regression model**. Hedonic price models analyse the price of a good depending on its multiple characteristics, by means of the price estimate implicit in each of them.

- **Flats**. These are the dwellings that are a part of a building with two or more floors, and which has a common access to all of them from a public road. So long as there are restricted areas and common areas, there is a special kind of co-property established as horizontal property.

- **Appraisal**. An appraisal is an estimation of the market value of a property, based on the different parameters determining it; in the case of dwellings, these parameters might be the surface area, the location, the age, etc. Most housing appraisal are carried out on request of a banking institution, for the purpose of the granting of a mortgage loan earmarked for the purchase of the dwelling, and they are carried out by valuation companies.

- **Value on drafting the deeds**. Drafting the deeds is confirmation by means of a public deed and in law of an issue or a act.

The value on drafting the deeds of a dwelling is that which appears as the value of the dwelling in the purchase and sale public deed, and is therefore the official price thereof.

- **Dwelling**. All structurally separate and independent venues that, given how they were constructed, reconstructed, transformed or adapted, are conceived to be inhabited by persons and form part of a building.

- **Second-hand dwelling**. Dwellings are classified as new or second-hand depending on the order of the transfer carried out. Thus, where there is more than one transfer on the public deed, the dwelling is deemed to be second-hand.

- **Free price dwelling**. This is a non state-subsidised dwelling.

- **New dwelling**. Dwellings are classified as new or second-hand depending on the order of the transfer carried out. Thus, when it is the first transfer on the purchase and sale deed, which is normally carried out by the developer or construction company on behalf of the first buyer, the dwelling is classified as new.

- **State-subsidised dwelling**. This is a dwelling which is subject to any type of subsidy for its construction, regardless of which body grants this, and where surface area and price limitations are taken into account. Those dwellings which have exceeded the time limit of the aforementioned subsidy are excluded, as are others which, although they have not yet exceeded it, appear with a construction value defined in a Ministerial Order by the Ministry of Economy and Tax. These last two considerations confer upon the dwelling the category of free price dwelling.

- **Single-family dwelling**. This is a dwelling located on an independent plot, accommodating a single family.

Annex II. Regression model

Specification of the regression model

The following describes the regression model that is used for calculating the prices estimated per square metre, used in compiling the HPI. For each quarter *q*, it is assumed that the price per square metre, *P*, of dwelling *i*, belonging to cell *c*, is:

$$l_{i,c}^{q} = \ln P_{i,c}^{q} = \mathbf{x}_{c}^{'}\boldsymbol{\beta}^{q} + \varepsilon_{i,c}^{q} \qquad (1)$$

where:

$\mathbf{x}_{c}^{'}$ is a vector of dimension *(1 x p)*, whose elements are equal to *0* or *1*, depending on the characteristics that define cell *c*, in terms of main effects and interactions,

β^{q} is a vector of *p* unknown parameters, of dimension (*p x 1*), and

$\varepsilon_{i,c}^{q}$ is the random component of the model, in quarter *q*.

Vector β^{q} defines the proportional effect on the price expected per metre squared of dwelling of *p* dichotomous variables included in $\mathbf{x}_{c}^{'}$, in quarter *q*. The *p* unknown parameters include the constant and the parameters of the dichotomous variables associated with the main effects and the interactions of the model.

For each *r* possible categories that has a main effect, the model includes (*r-1*) parameters. If the interaction has (*r x s*) possible combinations of values, the model will have *(r-1) x (s-1)* parameters. In total, the model in force in 2008 has 157 parameters.

The distortions $\varepsilon_{i,c}^{q}$ verify:

$$E[\varepsilon_{i,c}^{q}] = 0,\ \ Var[\varepsilon_{i,c}^{q}] = \sigma_{q}^{2},\ \ Cov[\varepsilon_{i,c}^{q}, \varepsilon_{j,d}^{q'}] = 0,\ \ \forall (q,i,c) \neq (q', j, d) \qquad (2)$$

Once the model is defined, and which will be in force for one year, the vector must be estimated $\boldsymbol{\beta}^{q}$ each quarter, with the information available. To this end, the model (1) is prepared in matrix notation, in the following manner:

$$\mathbf{L}^{q} = \mathbf{X}^{q}\,\boldsymbol{\beta}^{q} + \boldsymbol{\varepsilon}^{q} \qquad (3)$$

where:

$\mathbf{L}^{q}$ is a vector of dimension (n^{q} *x 1*) which contains the n^{q} elements $l_{i,c}^{q}$ of quarter *q*. That is, it contains as many rows as purchase and sale of dwellings that have taken place over quarter *q* (n^{q}),

$\mathbf{X}^q$ is a matrix of dimension ($n^q \times p$), whose elements are equal to *0* or *1*. In this matrix, each row represents a dwelling, and each column contains one of *p* characteristics that define said dwelling, in quarter *q*,

β^q is a vector of dimension (*p x 1*), which contains *p* unknown parameters in quarter *q*. This includes the constant and the parameters of the dichotomous variables associated with the main effects and the interactions of the model, and

$\boldsymbol{\varepsilon}^q$ is a vector of dimension (n^q *x 1*) , which contains the n^q random distortions of the model in quarter *q*. This distortion vector verifies:

$$E[\boldsymbol{\varepsilon}^q] = \mathbf{0}, \;\; Var[\boldsymbol{\varepsilon}^q] = \sigma_q^2 \, \mathbf{I}_{n^q \times n^q} \qquad (4)$$

The OLS (ordinary least squares) estimator [3] from β^q is:

$$\hat{\boldsymbol{\beta}}^q = (\mathbf{X}^{'q}\mathbf{X}^q)^{-1}\mathbf{X}^{'q}\mathbf{L}^q \qquad (5)$$

and its variance is:

$$Var[\hat{\boldsymbol{\beta}}^q] = \sigma_q^2(\mathbf{X}^{'q}\mathbf{X}^q)^{-1} = \mathbf{V}^q \qquad (6)$$

where the matrix V^q has dimension (*p x p*).

The vector of parameters $\hat{\beta}^q$ varies according to the data from each quarter, and is the fundamental element used for estimating the average price per cell.

Price estimation

In the compilation of the HPI, it is necessary to have, for each quarter, the estimated average price corresponding to each cell. This estimated price is obtained using the price from the formula (1); thus, the estimated price of cell *c*, in quarter *q*, is the following:

$$\hat{P}_c^q = \exp(\mathbf{x}_c^{'}\hat{\boldsymbol{\beta}}^q) \qquad (7)$$

The problem with this estimator, which has a simple expression, is that it has a high degree of bias. In order to correct this bias, the estimator proposed by El-Shaarawi and Viveros (1997) is used:

[3] The deduction from these results may be viewed, for example, in the texts by Peña (1993, 2002), Draper (1998) and Montgomery (2001)

$$\hat{P}_c^q = \exp\left\{ \mathbf{x}_c^{'} \hat{\boldsymbol{\beta}}^q - \frac{1}{2} \mathbf{x}_c^{'} \hat{\mathbf{V}}^q \mathbf{x}_c + \frac{1}{2} \hat{\varphi}^q \hat{\sigma}_q^2 \right\} \qquad (8)$$

where

$$\hat{\varphi}^q = 1 - \frac{\hat{\sigma}_q^2}{2(n^q - p)} - \frac{\hat{\sigma}_q^4}{3(n^q - p)^2} \qquad (9)$$

The estimator (8) substantially corrects the bias of the estimator (7), assuming the normality of the errors $\varepsilon_{i,\,c}^q$.

In order to obtain the estimation of the variance that appears in the above expressions, the residuals are defined $e_{i,\,c}^q$ as the difference between the naperian logarithms of the observed price and the estimated price, that is:

$$e_{i,c}^q = l_{i,c}^q - \mathbf{x}_c^{'} \hat{\boldsymbol{\beta}}^q \qquad (10)$$

The variance σ_q^2 is estimated with the average of the residual quadrants:

$$\hat{\sigma}_q^2 = \frac{1}{n^q - p} \sum_{c,\,i}^{n^q} \left(e_{i,c}^q \right)^2 \qquad (11)$$

Correction of heteroskedasticity

On applying the regression model to the data each quarter, the residuals present signs of heteroskedasticity for one of the variables included in the model, as well as for the set of observations that have imputed values. Therefore, a transformation must be carried out that makes the model homoskedastic.

In heteroskedastic models, the variance of the residuals is not constant, given that:

$$\text{var}[\boldsymbol{\varepsilon}^q] = \sigma_q^2 \left(\mathbf{W}^q \right)^{-1} \qquad (12)$$

where W^q a diagonal matrix with dimension (n^q x n^q) and all its positive elements.

Given that:

$$\text{var}\left(\left(\mathbf{W}^q \right)^{1/2} \boldsymbol{\varepsilon}^q \right) = \sigma_q^2 \, \mathbf{I}_{n^q \times n^q} \qquad (13)$$

the model can be made homoskedastic, pre-multiplying it by the matrix $\left(\mathrm{W}^q\right)^{1/2}$; in other words:

$$\left(\mathbf{W}^q\right)^{1/2}\mathbf{L}^q=\left(\mathbf{W}^q\right)^{1/2}\mathbf{X}^q\boldsymbol{\beta}^q+\left(\mathbf{W}^q\right)^{1/2}\boldsymbol{\varepsilon}^q \qquad (14)$$

The estimator $\hat{\boldsymbol{\beta}}^q$ that minimises the weighted sum of the squares of the errors is expressed as follows:

$$\hat{\boldsymbol{\beta}}^q=(\mathbf{X}^{q'}\mathbf{W}^q\mathbf{X}^q)^{-1}\mathbf{X}^{q'}\mathbf{W}^q\mathbf{L}^q \qquad (15)$$

and its variance is:

$$Var[\hat{\boldsymbol{\beta}}^q]=\sigma_q^2(\mathbf{X}^{q'}\mathbf{W}^q\mathbf{X}^q)^{-1}=\mathbf{V}^q \qquad (16)$$

The idea that justifies the introduction of the matrix W^q in the model is that the variance of the data is different for the different categories of a variable, the observations that belong to those categories with a lower variance are more reliable and must carry a greater weight in the weighted sum of squares of the errors than those with a greater variance (on average, the less the variance, the less they will deviate from the average value that we intend to estimate). Something similar occurs with the complete observations (without imputed values), which in general, have less variation that those in which it has been necessary to impute values.

The elements of the matrix W^q are determined using the analysis of the heteroskedasticity of the model. Thus, for the correction thereof, in the formula (8) of the estimated average price per cell, we must use the new expressions of $\hat{\boldsymbol{\beta}}^q$ and V^q, and the residual variance of the corrected model will be obtained using the weighted residuals:

$$e_{i,c}^q=\sqrt{w_i^q}\ \ (l_{i,c}^q-\mathbf{x}_c^{'}\hat{\boldsymbol{\beta}}^q) \qquad (17)$$

where W_i^q is the element (*i,i*) of the matrix W^q.

ASSIGNATION OF WEIGHTS OF HETEROSKEDASTICITY BY THE IMPUTATION OF VALUES

In the notary database, most of the variables that are directly or indirectly involved in the model are complete. However, when this is not the case, it is necessary to impute the values that are not informed.

As the variability of the residuals in the observations where the value of some of the explanatory variables of the model have been imputed is greater than in the set of whose that are complete in the data file, the complete observations are assigned a weight equal to one in the regression, whereas those with imputed values have a lower weight assigned.

For the calculation of these weights, we have used the average quadratic error (MCE): for the set of observations that have the value of a set of main effects U imputed, the corresponding weight is obtained as the quotient of the average quadratic error of the complete model, with all of the main effects (MCE_T^q) and the average quadratic error of the model that excludes the main effects and interactions associated with the set U of imputed variables (MCE_{T-U}^q). In order to calculate these terms, MCE_T^q and MCE_{T-U}^q , we use the set of complete observations C; that is, excluding all those observations from the quarter that have imputed the value of any of the main effects of the model.

As the complete model has a residual variation lower than that of the submodel that excludes one or more main effects (and their corresponding interactions), we can verify that:

$$0 \leq \lambda_u^q = \frac{MCE_T^q}{MCE_{T-U}^q} \leq 1 \qquad (18)$$

where:

MCE_T^q the average squares of the error of the model that includes all of the main effects and interactions applied to set *C* of observations without imputed values in quarter *q*, and

MCE_{T-U}^q the average of squares of the error of the model that excludes the main effects *U* in those in which some value has been imputed, applied to set *C* of observations without imputed values in quarter *q.*

t is logical to assume that those observations that have been subjected to an imputation procedure will have a greater variance of error (or a lesser weight in the adjustment of the model). In order to bear in mind this fact, we consider a heteroskedastic model of the type (12) where the weights W_i are defined as follows:

- If the *i-th* observation of quarter *q* has complete information, then $\mathrm{W}_i^{impu} = 1$.
- If the *i-th* observation of quarter *q* is incomplete and lacking the data corresponding to the set of explanatory variables *U*, then $\mathrm{W}_i^{impu} = \lambda_U$.

As many weights will be calculated λ_U as there are possible cases or combinations of main effects imputed given during the quarter. In the simplest case, only the valued of a main effect in the model will be imputed, and it will only be necessary to calculate a different weight of one.

CORRECTION OF HETEROSKEDASTICITY AMONG CATEGORIES

The analysis of the residuals of the previous weighted model can make it necessary to carry out a last correction of heteroskedasticity present in some of the explanatory variables. In order to carry out this correction, the steps to follow are explained below:

Whether they are C_1, C_2,...,C_u the U possible values of the variable on which the heteroskedasticity is going to be corrected:

1. The previously weighted model is adjusted.
2. The residuals of the previously weighted model $\hat{e}_i^q \;\; i = 1,...,n^q$.
3. The estimated variances of the residuals within each category are obtained:

$$S_r^2 = \frac{1}{n_r - 1} \sum_{i \in C_r} \left(\hat{e}_i - \bar{\hat{e}}_r \right)^2, \;\; n_r = card\left(C_r \right), \;\; \bar{\hat{e}}_r = \frac{1}{n_r} \sum_{i \in C_r} \hat{e}_i \qquad (19)$$

4. This defines

$$w_i^{cate} = \frac{\min\left[S_1^2,...,S_U^2 \right]}{S_1^2} \;\; \forall i \in C_1, \;\; ....., \;\; w_i^{cate} = \frac{\min\left[S_1^2,...,S_U^2 \right]}{S_U^2} \;\; \forall i \in C_U$$

The joint correction of heteroskedasticity is carried out with a weighted model, defining the weighting or weight of each observation as the production of the two coefficients calculated in the previous section, and in this one, in the following manner:

$$w_i^{hete} = w_i^{impu} \times w_i^{cate}$$

where w_i^{impu} the coefficient assigned to the *i-th* observation, bearing in mind the imputed values it has, and w_i^{cate} the coefficient assigned to the value or category of the variable that presents heteroskedasticity problems, in the *i-th* observation.

The matrix W^q of the homoskedastic model (14) is a diagonal matrix, of dimension $n^q \times n^q$, where the elements of the main diagonal are the coefficients w_i^{hete} .