INSTITUTO NACIONAL DE ESTADÍSTICA

Housing Price Index

Methodology

Madrid, 2009

INE. National Statistics Institute

Table of contents

1. Background

The present methodology explains the main characteristics and procedures used in the calculation of the Housing Price Index (HPI). This is an indicator conceived for the purpose of completing the information produced in Spain regarding the housing market.

Up until the publication of the HPI, the main statistics that provided information on the evolution of prices in the real estate market were those produced by the Ministry of Housing. In turn, the private sector, mainly appraisal companies and real estate agencies, also publish some information on housing prices.

Any of them can enable obtaining a perspective of the real estate market in Spain, but they do not show the sector situation from all possible scopes, thereby preventing the carrying out of a complete analysis thereof. This shortage of information is one of the main reasons that encouraged the National Statistics Institute (INE) to consider the implementation of a set of statistics covering the housing market in general, placing emphasis on hedonic regression procedures for processing quality adjustments, an aspect which is of vital importance in price indices.

The same concern has been present in the European Union (EU) countries for years. Due to this, in the year 2002 at the heart of the European Statistics Office (Eurostat), a Study Group was created to face the problem of measuring the prices of housing purchases. Spain has been a part of this Group since its beginning, currently comprised of most countries of the EU.

The work carried out within the Study Group has been performed in three phases, in each one of them broadening the objectives pursued, which have gone from the study of the market situation in each country (available information sources and valid calculation methods) in the first phase, to the periodical production of indices enabling measuring the evolution of new and second-hand housing prices, the objective of the third stage, which ends in the middle of the year 2009.

With this initial situation, it is essential to achieve the implementation of statistics intended to ascertain the time evolution of the purchase prices of housing in Spain, meeting the current information need and enabling establishing international comparisons.

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

In turn, in 2005, the INE created an Interministerial and Bank of Spain Working Group, with representatives from the Government Presidency, the Bank of Spain and the Ministries of Economy and the Treasury, Justice and Housing. This Working Group presented and analysed all of the elements that must comprise an indicator of the evolution of housing prices, with the basic characteristics that enable the required international comparison.

This document explains the procedure followed in the years prior to implementing the HPI, as well as the most relevant methodological characteristics defining it.

Particular attention has been paid to the method developed for processing the information obtained, based on regression models, which differentiates these statistics from those that had been compiled from the public sector up until now.

2. Background. Eurostat Study Group

The creation of the Study Group on owned housing 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 work on the work project followed in order to obtain the HPI.

The **first phase** of the project (between the years 2002 and 2004) for this group, comprising Germany, Spain, Finland, Poland and the United Kingdom, had the purpose of studying the situation of the real estate market in each one of the countries, locating the main sources of information available and proposing a statistical calculation procedure that would enable making comparisons of 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 surveying 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 the survey values, and for which work was performed with average survey values made by the most important surveying 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 treatment 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 (standardisation). This is the main reason why the project for implementing the HPI must have as a reference those general guidelines set out by Eurostat for the harmonised indicator.

Moreover, each country had to face the job 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, requiring that the prices used to

calculated the indices be actually paid prices of the merchantings 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 merchanting process. In addition to the two sources previously mentioned (surveys and real estate developments), there are other agents that provide information regarding housing: drafting of deeds, mortgage loans, Land registers and 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 drafting of deeds values provide the best information for monitoring 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 the survey 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 surveyed are subject to merchanting, 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. From a practical point of view, given that this information is from the notary and property registers, the availability - over time and in form - is the same.

The data from the Cadastre has also been dismissed, because the real estate 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. Moreover, the complexity and cost of said operation make it unfeasible.

Of all of these sources, that which was eventually selected for the calculation of the HPI is the notary register, which contains, among other data, the official prices for all of the merchantings occurring in Spanish territory, and correspond to the value of the public 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 following lists the most noteworthy characteristics presented by the values of the public deeds:

- **Availability**. At the beginning of 2005, the General Notary Council and the INE signed a Partnership Agreement for conducting statistical operations regarding real estate, which enables having the information necessary for calculating the HPI (data relating to housing transactions) with the periodicity and terms that are adequate for the statistical operation in question.

- **Punctuality**. Each month, the INE receives the information from the transfers of real estate property taking place in Spain, with a lag of approximately six weeks from the end of the reference quarter of the data. This term has been agreed on, for the purpose of including more than 90% of the total transactions taking place in the reference quarter.

- **Present**. The transactions are part of the calculation of the index the same quarter in which they take place.

- **Coverage**. This includes information regarding the merchantings carried out in the country throughout the quarter, of both new dwellings and second-hand dwellings, and regardless of the form of payment.

- **Content**. The notary database contains detailed information that enables establishing very specific dwelling typologies. The variables received refer to characteristics of the dwelling, as well as of the buyer; among the former, worth noting are those relating to the size of the property and its location.

On the other hand, and 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 o the sample of dwellings that are 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.

Lastly, in December 2007, the third phase of the Study Group began, in which 26 EU countries participated, and among whose objectives are the periodical production of indices and the improvement of its calculation methodology. Moreover, the index of the costs associated with housing acquisition will continue to be calculated, and an index of large repairs will begin to be calculated.

3. Objectives

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

Not included within its scope is the measurement of price levels. Therefore, spatial comparisons cannot be established for the price levels, whereas they can for the price evolutions.

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 o 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, the HPI 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 HICPs of the EU.

4. Research scopes

Units of analysis

Bearing in mind that the objective of the HPI is to measure the evolution of the level of merchanting prices of free price housing in Spain, the unit for analysis is free price housing.

Population scope

The HPI population or reference stratum includes the entire population (individuals), resident both inside and outside of Spain, who have acquired a dwelling during the reference period. Those purchases made by companies 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 whole of Spain.

Time scope

The HPI is carried out quarterly, which enables estimating the evolution of prices between consecutive quarters, that accumulated over the year, as well as the annual evolution.

5. Information source. Variables used

As mentioned previously, the information used for calculating the HPI is from the General Council of Notaries, with which the INE signed a partnership agreement intended to enable the use of the data from the 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.

Every month, ANCERT provides the information from the transfers of property occurring in Spain, in addition to the subsequent updates in which new observations or modifications are included of those previously sent.

The HPI is a short-term, quarterly, index, and involved in its compilation is the latest data that the INE has at the time of calculation, which represents, on average, more than 90% of the total transfers carried out each quarter. This coverage more than meets the requirements of this indicator, bearing in mind the punctuality and opportunity criteria expected of statistics of this nature. Nonetheless, it has been verified that the information received after the indicator has been calculated has a negligible effect on the final results.

The information submitted monthly by ANCERT includes the following variables:

- **Dwelling location variables:**
  - Autonomous Community.
  - Province.
  - Municipality.
  - Postal code.
  - Type, name and number of the road.
  - Duplicate, block, stairwell, floor and door.
- **Time variable indicating the day on which the dwelling is transferred:**
  - Date of authorisation.
- **Legal act codes**. The transactions corresponding to the following legal acts are received:
  - Merchanting of real estate property.
  - Cooperative allocation of housing to its members.
  - Commune allocation of a real estate development to a commune member in the 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.

- **Variables relating to features of the dwelling:**

  - Type of real estate property. This indicates the type of urban property; a distinction is made between dwellings, parking spaces, storage rooms and plots of land.

  - Reference in the land registry of the property, or reason for absence or for not being able to obtain this, where appropriate.

  - Type of dwelling, distinguishing between flat and single-family dwelling.

  - 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 merchanting deed, 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, where there is more than one transfer on the public deed, the dwelling is deemed to be second-hand.

  - Surface area built in m^2.

  - Variables indicating whether the parking space and storage room are included in the price.

  - Variable indicating whether the dwelling is free-price or official protection.

- **Variables relating to the buyer:**

  - Type of person. Transaction are received from individuals, whilst those made by legal entities (companies) or financial institutions are excluded.

  - Country, province and municipality of residence of the buyer.

  The variables relating to the buyer enabled the adaptation to the population coverage of the HPI, identical to that of the HICP, which must include those purchases made by individuals in the national territory (country), regardless of their nationality.

In addition to the information received from the General Council of Notaries, other variables have been generated, using the information from other sources, which enable increasing the explicative capacity of the regression model used in estimating prices. These variables are the following:

- **Cluster of provinces**. Grouping of the 52 provinces into 6 groups, through the application of a cluster analysis, using the classification variable of the average annual value of the mortgage, obtained from the INE Mortgage Statistics considered solely for urban properties, and more specifically for dwellings.

- **Size of the municipality**. Classification of the municipalities, distinguishing between provincial capitals and large non-capital municipalities (with more than 100,000 inhabitants), medium-sized municipalities (with 10,000 to 100,000 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**. This is obtained from the information on the number of annual overnight stays in each type of tourist establishment, provided by the Hotel Occupancy, Holiday Dwelling Occupancy and Rural Tourism Accommodation Occupancy (INE) surveys.

- **Type of environment**. Classification of the postal codes in 14 categories, using the information from the 2001 Population and Housing Census regarding the average socio-economic condition and the average level of problems, considering problems to be aspects such as crime, noise level or dirtiness in the area, poor communications, a lack of green (park) areas, etc. This has also considered the actual price, per metre square, by postal code, in a prior annual period.

6. 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 over time 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 measurement of the evolution thereof would not be correct.

A possible solution is the creation of dwelling strata that group those with similar characteristics, and to follow the average price of each stratum instead of the price of each dwelling individually, to subsequently obtain the joint evolution, weighting each one of the strata. In this sense, it is appropriate to create small strata; since the more defined the dwelling typology is, the better the adjustment will be due to the change in 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 hedonic regression models that enable the quarterly estimation of prices by stratum, regardless of whether there are 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.

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.

6.1 Prices

As stated previously, the prices per metre squared are one of the basic elements in the calculation of this indicator. However, given the diversity 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. This process uses the original information regarding purchases-sales of dwellings for each quarter, and to which a regression model is applied.

The validation and filtering process of the data for the calculation of the HPI is carried out in two stages. On the one hand, before the application of the regression model, the information is analysed and the atypical values found are detected and corrected. On the other hand, the regression model itself detects the atypical observations, which are subsequently corrected, based on the joint information provided by all of the explicative variables of the model.

Thus, the initial regression model is applied to the file of data resulting from the first phase of validation and filtering, and in this way, the atypical data from the model is detected and eliminated.

For the years 2007 and 2008, approximately 52,000 different housing typologies have been constructed, and which have been decided by observing the physical characteristics and location of the target dwellings of merchanting over the course of three years: 2005, 2006 and 2007. The prices of each typology are estimated quarterly, using the merchantings of dwellings carried out during the quarter, regardless of the number of quarterly transactions for each typology.

The model therefore enables obtaining an estimated price for all housing typologies, regardless of the number of transactions carried out that quarter in that stratum, using the information supplied by the quarterly sample of merchantings. Finally, the average quarterly price for each typology is obtained by multiplying the parameter vector (which varies quarterly) by the vector of the characteristics defining each housing typology. Said parameters include the implicit prices for these characteristics.

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

6.2 Regression model

In the regression model applied for the calculation of the HPI, 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. Each possible combination of the values of these variables comprises what is known as a cell.

This model has the dependent variable of the neperian logarithm of the price per metre squared of the dwelling; it thus includes the non-linear relationship between the price and the remaining variables, and also enables a simple interpretation of the coefficients or parameters of the model.

The model that has been used for estimating prices in the years 2007 and 2008 includes the following main effects or explicative variables:

- New/Second-hand.
- Type of dwelling (distinguishing between flats and houses).
- Parking space (Yes/No).
- Storage room (Yes/No).
- Cooperative (Year/No).
- Surface area (in brackets).
- Cluster of provinces.
- Size of the municipality.
- Tourist-oriented municipality (Yes/No).
- Type of environment.

The first five are dichotomous variables that are obtained directly from the data file of the notaries; the last four are created using the information from other sources, as described in section 5. Lastly, the surface area variable has been formed, classifying the values of the *surface area built of the dwelling* numerical variable:

1. $< 40\ m^2$
2. $[40, 60)\ m^2$
3. $[60, 75)\ m^2$
4. $[75, 90)\ m^2$
5. $[90, 105)\ m^2$
6. $[105, 120)\ m^2$
7. $[120, 150)\ m^2$
8. $[150, 180)\ m^2$
9. $[180, 240)\ m^2$
10. $\geq 240\ m^2$

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 50 in each one of the combinations or pairs of possible values of the interaction.

The model used for the data from the year 2008 consists of 9 interactions, with together with the main effects come to a total of 157 parameters to estimate each quarter.

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

- updating with the latest available information, of the sources used in its compilation, the cluster variables of provinces, size of the municipality, tourism-oriented municipality and type of environment, which can make the number of categories thereof vary;

- in parallel, there may be an incorporation, in the model, of new explicative variables from the database of the notaries or created using complementary information.

Annex II includes the specification of the regression model, the correction of heteroskedasticity and the estimation of the prices per cell.

6.3 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) whose composition may change from year to year, 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.

The weighting structure is obtained with the information referring to the last three available years. Two objectives are thereby sought:

- To guarantee the stability of the weighting structure, since if it changes significantly from one year to the next, the annual rates could be affected.

- To improve the representativeness of the index. The more years are involved in the calculation, the greater the variety of typologies or strata of dwellings represented, and therefore, the better the adjustment due to changes in quality and composition.

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. It will be possible to carry this out, as the HPI is a linked Laspeyres index.

As mentioned above, the weighting of each stratum represents the relationship between the expenditure that households make on the purchase of dwellings of the same typology, that is, belonging to the same stratum, and the total expenditure made on the purchase of housing. The estimated expenditure has been used for the calculation of the weightings, calculated by multiplying the price per m^2 of the dwellings included in stratum *e* for the fourth quarter of the year prior to the current year ($\hat{P}_e^{4,\,a-1}$), per m^2 sold of dwellings belonging to said stratum during the reference period:

$$W_e^a = \frac{Q_e^{(a-1,a-2,a-3)} \times \hat{P}_e^{4,\,a-1}}{\sum_{\forall e} Q_e^{(a-1,a-2,a-3)} \times \hat{P}_e^{4,\,a-1}} \qquad a \geq 2008$$

where,

$Q_e^{(a-1,a-2,a-3)}$ the average annual amount of metres squared of dwellings belonging to stratum *e*, sold during the reference period of weightings (*a-1*, *a-2*, *a-3*),

W_e^a the weighting of stratum *e*, in so much per one, valid during year *a*.

The reason why the expenditure on housing is calculated using the prices estimated by the regression model for the last quarter of the previous year, is to correct the discrepancy between the reference period of the amounts (years *a-3*, *a-2*, *a-1*) and the reference period of the prices (the 4th quarter of *a-1*). In this way, using the latest information available regarding prices, the expenditure of each typology of dwellings is updated to the last quarter of the previous year, which is also the reference quarter for the prices.

The fact of using estimated prices rather than collected prices is due to the fact that, during the reference period (the last quarter of year *a-1*), it is possible that dwellings from all of the typologies have not been sold, and therefore, information is not available for transacted prices for all of the strata.

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:

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

6.4 Calculation of indices

The general formula used in the calculation of the HPI is a linked Laspeyres index, analogous to that used in the Consumer Price Index (CPI). In the case of the HPI, since it is a quarterly indicator, the period used for linking is the fourth quarter of each year, and not the month of December.

The use of linked indices enables the annual updating of the weightings, as well as the possibility of making methodological changes (review of the model, inclusion of new strata, etc.), 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 chain-linked index, three reference periods are defined:

- **Reference period of the base index or period.** This the reference period in which all indices are made equal to 100. It usually involves an annual period. In the HPI, the base year is 2007, and all the indices published since then will have as a reference the aforementioned period.

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

Each year, the weightings of the HPI are updated with the latest available information regarding the merchantings of dwellings taking place in the last three available years, updated to prices for the 4th quarter of the year prior to the current year. Consequently, the reference period of the weightings will vary each year.

In the year 2008, the reference period of the weightings will be that comprising the years 2005, 2006 and 2007, updated to prices from the 4th quarter of 2007.

- **Reference period of the prices**. This is the period with whose prices the current prices are compared; in other words, the period chosen for calculating the simple indices. This is the fourth quarter of the year immediately preceding the current year, and therefore, it will vary each year.

The following expresses the general calculation formula of the basic indices and the aggregated indices, as well as the general calculation system of the chain-linked indices.

BASIC INDICES

A basic index 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 (groupings) are called basic indices. In the HPI, the basic aggregation is the stratum that includes a single typology of dwelling (housing).

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:

$$ {}_{(4,\,a-1)}I_e^{q,a} = \frac{\hat{P}_e^{q,a}}{\hat{P}_e^{4,\,a-1}} \times 100 \quad a \geq 2008 $$

where,

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

$\hat{P}_e^{4,a-1}$ the price estimated per metre squared for those dwellings in the stratum, 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.

GROUPED 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:

$$ {}_{(4,\,a-1)}I_A^{q,a} = \sum_{c \in A} W_e^a \times {}_{(4,\,a-1)}I_e^{q,a} \quad a \geq 2008 $$

where,

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

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

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

INDICES IN BASE 2007

The indices in base 2007 are those that are published and obtained linking the indices referring to the fourth quarter of the previous year, according to the following expression:

$$ {}_{07}I_A^{q,\,a} = {}_{07}I_A^{4,\,(a-1)} \times \left(\frac{{}_{4,(a-1)}I_A^{q,\,a}}{100} \right) = \quad a \geq 2008 $$

$$ = {}_{07}I_A^{4,\,07} \times \left(\frac{{}_{(4,07)}I_A^{4,\,08}}{100} \right) \times \ldots \times \left(\frac{{}_{(4,\,a-2)}I_A^{(4,\,a-1)}}{100} \right) \times \left(\frac{{}_{(4,\,a-1)}I_A^{q,a}}{100} \right) $$

[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* will calculated the denominators of the basic indices of the four quarters of year *a+1*.

For base year 2007, the indices are defined as set out below:

$$ {}_{07}I_A^{q,07} = \frac{\sum_{e\in A} Q_e^{(05,06,07)} \times \hat{P}_e^{q,07}}{\sum_{e\in A} Q_e^{(05,06,07)} \times \bar{\hat{P}}_e^{07}} \times 100 $$

with $\bar{\hat{P}}_e^{07} = \frac{1}{4}\sum_{q=1}^{4} \hat{P}_e^{q,07}$; therefore: $\frac{1}{4}\sum_{q=1}^{4} I_e^{q,07} = 100$

Using the indices in base 2007, the quarterly variation rates are obtained, the accumulated (or year-to-date) variation rate and the annual variation rate. The quarterly and accumulated variation rates can also be obtained using the indices referring to the 4th quarter, as may be viewed below.

6.5 Calculation of variation rates

QUARTERLY VARIATION RATE

The quarterly variation rate of an index in a period (q, a) is calculated as the quotient between the index for current quarter q, and the index for previous quarter q-1, both in base 2007, according to the following formula:

$$ \Delta^{qa/(q-1)a} = \left(\frac{{}_{07}I^{qa}}{{}_{07}I^{(q-1)a}} - 1 \right) \times 100 $$

where:

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

${}_{07}I^{qa}$ the index of quarter *q* in year *a*, in base 2007, and

${}_{07}I^{(q-1)a}$ the index of quarter *q-1* of year *a*, in base 2007.

These rates can also be calculated with the indices referring to the fourth quarter of the immediately preceding year, obtaining the same result as with the indices in base 2007.

$$\Delta^{qa/(q-1)a} = \left(\frac{{}_{07}I^{qa}}{{}_{07}I^{(q-1)a}} - 1 \right) \times 100 = \left(\frac{{}_{07}I^{4(a-1)} \times \frac{{}_{(4,a-1)}I^{qa}}{100}}{{}_{07}I^{4(a-1)} \times \frac{{}_{(4,a-1)}I^{(q-1)a}}{100}} - 1 \right) \times 100 =$$

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

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 2007:

$$\Delta^{qa/4(a-1)} = \left(\frac{{}_{07}I^{qa}}{{}_{07}I^{4(a-1)}} - 1 \right) \times 100$$

where:

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

${}_{07}I^{qa}$ the index in base 2007, in quarter *q* of year *a*, and

${}_{07}I^{4(a-1)}$ the index in base 2007, in the fourth quarter of year *a-1*.

As with the quarterly rates, the accumulated rates can also be calculated with the indices referring to the fourth quarter of the previous year, obtaining the same result as with the indices in base 2007.

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 2007:

$$\Delta^{qa/q(a-1)} = \left(\frac{{}_{07}I^{qa}}{{}_{07}I^{q(a-1)}} - 1 \right) \times 100$$

where:

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

${}_{07}I^{qa}$ the index in base 2007, in quarter *q* of year *a*, and

${}_{07}I^{q(a-1)}$ the index in base 2007, in quarter *q* of year *a-1.*

6.6 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 did not vary 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 s given stratum *e*, is the product of its quarterly variation rate of prices, $\Delta_e^{qa/(q-1)a}$, by its weighting, W_e^a, and by the quotient between the index for the stratum and the general index 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 basket of dwelling typologies of the HPI is the quarterly variation of the general index.

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

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

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

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

ACCUMULATED REPERCUSSIONS (for the year-to-date)

The accumulated repercussion, or the repercussion 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:

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

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

where:

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

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:

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

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

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

7. Dissemination

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, broken down information is provided regarding new housing and second-hand housing, on a national level.

In the coming years, the possibility will be evaluated of providing another type of more broken-down information, according to the type of housing or the characteristics thereof.

The following data will be published quarterly:

- indices 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).

Annex I. Glossary of terms

- **ANCERT**. Notarial Certification Agency, previously known as the Notarial Institute for Information Technologies (INTI). This is a company in the General Council of Notaries, created for the technological modernisation of the Spanish notary system.

- **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 (association)**. 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.

- **Survey**. A survey 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 surveys are carried out for 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 surveying 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 public merchanting 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 merchanting 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 (house)**. This is a dwelling located on an independent plot, accommodating a single family.

Annex II. Regression model

Specification of the regression model

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

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

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 a main effect has, 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:

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

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:

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

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 merchantings 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:

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

The MCO (minimal ordinary quadrant) estimator [2] from β^q is:

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

and its variation is:

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

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:

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

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

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:

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

where

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

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 variation that appears in the above expressions, the residuals are defined $e_{i,\,c}^q$ as the difference between the neperian logarithms of the observed price and the estimated price, that is:

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

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

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

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 variation of the residuals is not constant, given that:

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

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

Given that:

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

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

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

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

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

and its variation is:

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

The idea that justifies the introduction of the matrix W^q in the model is that the variation of the data is different for the different categories of a variable, the observations that belong to those categories with less variation are more reliable and must carry a greater weight in the weighted sum of squares of the errors than those with a greater variation (on average, the less the variation, 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 variation of the corrected model will be obtained using the weighted residuals:

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

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 sub-model 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.*

It is logical to assume that those observations that have been subjected to an imputation procedure will have a greater variation 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 $W_i^{impu} = 1$.

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

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_{or} 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 variations of the residuals within each category are obtained:

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

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}.