Using administrative data sources to develop real estate price statistics: The case of Portugal

Rui Evangelista

Statistics Portugal, National Accounts Department [1,2]

Abstract:

The most recent economic and financial crises highlighted the lack of important information to monitor the performance of the housing market, whose imbalances could, through wealth and other effects, impact significantly on economic growth. This information gap has prompted statistical institutes across Europe to develop strategies leading to the improvement and production of real estate price statistics. This paper presents the work of Statistics Portugal in this area, which involved the compilation of indicators based on prices information taken at different stages of the buying and selling process. This has enabled the comparison of different indexes and to an assessment of the possible impact of data source choice on house price inflation measurement. The chosen approach, which was based on the exploration of administrative data rather than on the implementation of new (and costly) surveys, has proven to be effective in providing indicators that can meet European Union legal requirements on house price indexes and satisfy the need for higher level policy indicators for macroeconomic surveillance.

Keywords:

Accessing Administrative Data, Policy Indicators, Hedonic Regression, House Price Indexes

JEL Classification:

C81, R31, E31, C43

[1] The views expressed in this paper are solely those of the author and do not necessarily reflect the position of Statistics Portugal or of any other institution mentioned in it.

[2] Contact email: rui.evangelista@ine.pt.

Introduction

The most recent economic and financial crises highlighted the lack of some important information on the housing market. This information gap has prompted statistical institutes across Europe to develop strategies to improve the already existing house price statistics and to start the production of new official real estate market indicators.

In the case of Portugal, data on transaction prices were generally not available for price index compilers, and the only indication about the evolution of residential housing prices was provided by a private producer, which uses asking prices collected from a real estate portal to compile a price index [1]. Statistics Portugal's chosen strategy to fill this information gap was based on the exploration of bank appraisals and fiscal administrative data sources rather than in the design of new (and costly) surveys[3]. The application of this "double data source approach" has proven to be effective in producing house price indexes (HPIs) and other real estate statistics without jeopardising the need to comply with the new European legal requirements in this area[4].

This paper provides an account of the work that was carried out in developing new HPIs and other real estate statistics for Portugal using administrative data. Of possible interest to future index compilers is the fact that, since produced HPIs are based on prices information collected at different stages of the buying and selling process, the work that was developed by Statistics Portugal provides a rare account of the likely impact of data source choice on house price inflation measurement.

The paper is divided into four sections. The first two sections present the data and the methodology used in the compilation of the new real estate statistics. The third section provides a summary of the main results obtained for the period starting in the first quarter of 2009 and ending in the last quarter of 2013 (1Q2009 - 4Q2013). Finally, some conclusions and final remarks are presented at the end of this paper.

[3] Administrative data simply refers to information collected by public or private institutions for administrative (not statistical) purposes.

[4] These include Regulation No 93/2013, which defines the third quarter of 2012 as the start for the regular production and transmission of HPIs to Eurostat. Another obligation stems from Regulation No 1176/2011, which includes the HPI in the scoreboard of indicators that is used for the early detection of macroeconomic imbalances in member states of the European Union.

1. Description of available administrative data

The first source of information – bank appraisals-, is generated through a monthly survey covering the most important commercial banks providing mortgage credit in Portugal[5]. This information refers to information that is collected before any actual purchase takes place. Although with a national coverage, transaction prices are out of the scope of the survey. Moreover, the survey lacks variables describing the quality of appraised properties and it is subject to criticism as it may be argued that the dwellings that are purchased in cash are not well represented by those that are typically purchased with mortgage credit. For the 1Q2009-4Q2013 period, the number of bank appraisals available for the compilation of real estate statistics exceeds the 326 thousand (an average of more than 16.3 thousand observations per quarter).

The second source of information – fiscal administrative data -, is derived from the combination of two different tax data sources, which are transmitted to Statistics Portugal on a monthly basis [6].

The records of the Municipal Tax on Real Estate Transfer (IMT), which provide information on transaction prices, constitute the first tax data source. The IMT is a tax levied on property transfers, which is calculated based on the declared value of the transaction or on the fiscal appraisal value of the property, depending on which is higher[7]. The IMT is paid by the buyer immediately before the property changes hands[8]. Available IMT data go as far back as the beginning of 2009.

The records of the Local Property Tax (IMI), which provide information on the characteristics of transacted dwellings, compose the second tax data source. The IMI is a municipal tax levied on the value of the property. The value relevant for IMI purposes is an appraisal of the value of the property (and not its market price), which is based on

[5] In 2011, sampled banks accounted for nearly 95% of all mortgage loans provided to private households by financial institutions.

[6] The access to fiscal administrative information is supported by data transmission agreements, which have been signed between fiscal authorities and Statistics Portugal. The development of real estate statistics using these data benefited from the existence of a good institutional and technical collaboration between the tax authorities and Statistics Portugal.

[7] It is generally accepted that under this system, which takes into account declared transaction prices and (fiscal) appraised values, the incentive to underdeclare is reduced and that the IMT generates information on transaction values that are the same (or very close) to real market transaction values.

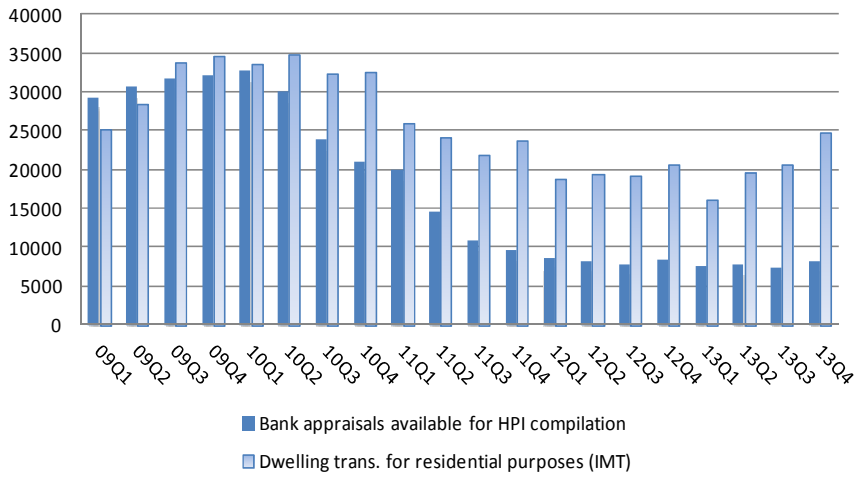
[8] A proof of payment of the IMT has to be shown before any dwelling transaction takes place. Typically, the payment of the IMT is done some days before the moment the transference of the property takes place.

an evaluation of its attributes (e.g., area, quality of location, etc). At present, Statistics Portugal has information about the characteristics of nearly 5.2 million dwellings[9].

For the compilation of new statistics, the IMT and IMI data sources were merged using a property (cadastral) register identification number, which is used as the key variable for the match of dwelling prices and characteristics. The end product of this matching and data cleaning process was a unique dataset with information characterizing both the prices and attributes of residential property transactions in Portugal. The number of transactions available for the 1Q2009-4Q2013 period exceeds the 450 thousand (i.e., an average of 22.5 thousand observations per quarter).

The next chart illustrates the number of bank appraisals and transactions available for the 1Q2009-4Q2013 period.

Chart 1: Number of bank appraisals and dwelling transactions



The chart represents the evolution of the residential property market in the last years. The number of bank appraisals only outscores those of transactions in the first two quarters of 2009, when mortgage credit was more abundant in the economy. During the 2Q2010 – 3Q2012 period, bank appraisals drop considerably and, with the exception of two quarters, at a faster rate than transactions. Finally, from 4Q2011 onwards, the

[9] According to the last (2011) population and housing Census, Portugal had a stock of 5.8 million conventional dwellings for residential purposes. The difference of around 600 thousand dwellings between the Census and the IMI figures can be explained (at least partly) by the fact that Statistics Portugal did not receive fiscal appraisals that were carried out from the start of the IMI and IMT tax system (December 2003) and the date in which the first data transmission agreement started to produce effects (beginning of 2005).

number of bank appraisals seems to have stabilized; accounting for far less than half of the number of recorded transactions in the last quarters of the series[10].

2. Methodologies used in the compilation of real estate statistics

2.1. Appraisals-based HPI

The HPI based on bank appraisals is compiled using a simple stratification approach. The stratification is based on four attributes considered to have a decisive influence on the formation of dwelling prices. The first one is "location" (seven NUTS II regions), the second attribute is "dimension" (two categories, defined on the basis of a threshold value for the number of rooms), the third one is "dwelling type" (i.e., whether it is a house or an apartment) and the last one is "occupancy status" (i.e., whether the dwelling is new or existing). The application of this scheme yields 56 elementary indexes, which are compiled using the geometric mean formula.

2.2.Transactions-based HPI

The transactions-based HPI is a hedonic price index that is derived from the application of the adjacent time dummy approach[11]. For all $q = Q-1, Q$ and $i = 1,...,n(q)$, this approach can be described as follows:

$$\eta_{i,q} = a + \sum_{K=1}^{K} \beta_k X_{i,q;k} + \theta_Q D_{i,q;Q} + \varepsilon_{i,q} \qquad (1)$$

where,

[10] The transactions data exhibits a clear seasonal pattern in the first and last quarters of each year. These quarters have a tendency to show the lowest and highest sales frequencies within each year, respectively. This is explained by the fact that buyers and sellers anticipate changes in house taxation that are implemented at the beginning of each year.

[11] A hedonic price index is simply an index that uses information taken from (what was dubbed as) a hedonic function [2]. The hypothesis underlying the building of hedonic functions is based on the reasoning that goods, such as houses, are purchased as bundles of characteristics and that price differences, found among similar varieties of the same good, are simply explained by the different quantities of quality attributes characterizing each one of these varieties [3]. The aspects underlying the estimation of hedonic functions have been the focus of many studies. A classical reference is [4] and a more recent reference is [2]. This method is recommended by international technical manuals in the construction of constant-quality house price indexes (see, for instance,[5]).

$\eta_{i,q}$, is the natural logarithm of the price level of the *ith* dwelling transaction in quarter *q*;

$X_{i,q}$, stands for the value of the *kth* characteristic of the *ith* transacted dwelling in quarter *q*;

D_Q, is the temporal indicator of quarter *Q*, which is defined as:

$$\text{for all } q = Q-1, Q \text{ and all } i = 1,\ldots,n(q), \quad D_{i,q;Q} = \begin{cases} 1, \text{ if } q = Q, \text{ and} \\ 0, \text{ if otherwise.} \end{cases}$$

θ_Q, is the parameter associated to the temporal indicator of quarter *Q*; and

$\varepsilon_{i,q}$, corresponds to an error term.

The value of the price index in quarter Q is obtained by the exponentiation of θ [12]. The parameters of (1) are estimated by ordinary least squares (OLS) for existing apartments, existing houses, new apartments and new houses[13].

In the specification process, special attention was given to location, area and age effects[14]. Robust statistics and individual and joint parameter significance tests were applied in the specification and estimation process[15]. Moreover, the Variance Inflation Factor (VIF) was applied to detect excess of collinearity and a (robust) version of the Ramsey ([6]) Reset Test was used to detect possible functional form problems. In order to investigate the influence of outliers in OLS results, least square absolute deviation (LAD) estimates were also computed and compared against OLS hedonic indexes.

2.3. House sales indicator

The house sales indicator is based on IMT data, which were restricted to reflect transactions of residential properties only. Agricultural land, commercial and non-arms length transactions (e.g., inherited dwellings) were excluded from the scope of the indicator. As in the transactions-based HPI, transactions of parts of dwellings were excluded from the calculations of the indicator. The total number of house sales can be

---

[12] For the estimation of $\exp(\theta)$, [7] proposes the use of $\exp\left(\hat{\theta} - 0.5\hat{\delta}^2\right)$, where $\hat{\delta}^2$ is an estimate of the variance of θ, instead of $\exp\left(\hat{\theta}\right)$. The results given by the two estimators are not expected to differ much in large samples and this correction is usually disregarded in hedonic time dummy studies.

[13] The separation between new and existing dwellings was directly imposed by Commission Regulation (EC) No 93/2013, which requires the provision of HPIs for these two subdivisions. The use of type of dwelling as a further stratification variable was based on the analysis of the data, which suggested that the apartments and house market segments should be dealt separately.

[14] The importance of micro and macro location levels in hedonic models is expressed in [8]. An example of a study on the relationship between age and market price of properties is given in [9].

[15] The presence of heteroskedasticity was checked through (a robust version of) the Breush-Pagan test. The homoskedasticity assumption was rejected in all the tests.

broken down into apartments and houses, NUTS II region and for the new and existing dwellings category.

3. Results

Using the above-mentioned administrative data sources, it was possible to compile appraisals- and transactions-based HPIs. As differences between these two indexes (HPI_BankA and HPI_Hed, respectively) may stem not only from used data sources but also as a consequence of the application of different methodologies, it was also compiled a price index (HPI_Strat) using fiscal administrative data, which used the same stratification approach applied in the appraisals-based HPI. For comparison purposes, the results of these indexes were also compared with an unadjusted HPI (HPI_raw)[16] and the asking-based house price indicator (HPI_Ci), which is provided by a private compiler[17].

The next page depicts two charts that provide information about index levels and year-on-year rates of change of the above-mentioned HPIs (100 = 1Q2009). Below each chart it is possible to see a table with the period's mean and standard deviation statistics. The house sales indicator (N_trans) is also presented in chart 2. This chart is provided with a double scale in which the left-hand side scale refers to the HPIs and the right-hand side scale refers to the house sales indicator.

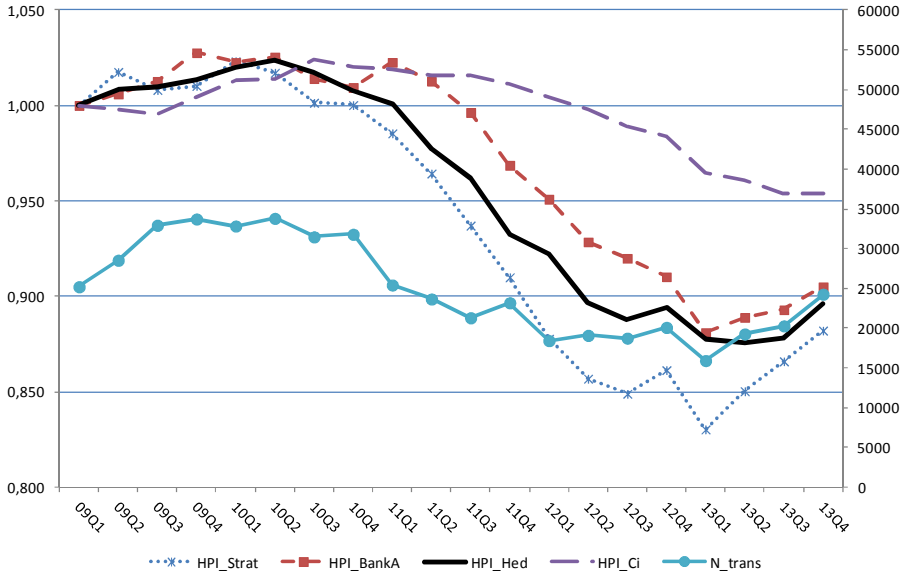
There are a few issues that are worthwhile to point out from the comparison analysis of the charts and index results. Firstly, despite the sharp drop in the number of bank appraisals (see chart 1), appraisals-based HPI seem to mimic the evolution of the transactions-based indicator reasonably well[18]. Secondly, it is important to highlight the fact that the asking prices HPI seems to lag behind both appraisals- and transactions-based HPIs. For the 1Q2010-4Q2013 period, the contemporary correlation between the

[16] The HPI_Raw used the same dataset and the four basic strata that were applied in the construction of the HPI_Hed. However, the former index does not apply the hedonic method and is simply based on the application of the geometric mean formula in each stratum.

[17] There is no information available about the geographical coverage of this index, which is compiled monthly and is based on the estimation of a hedonic function. More information can be found in [1].

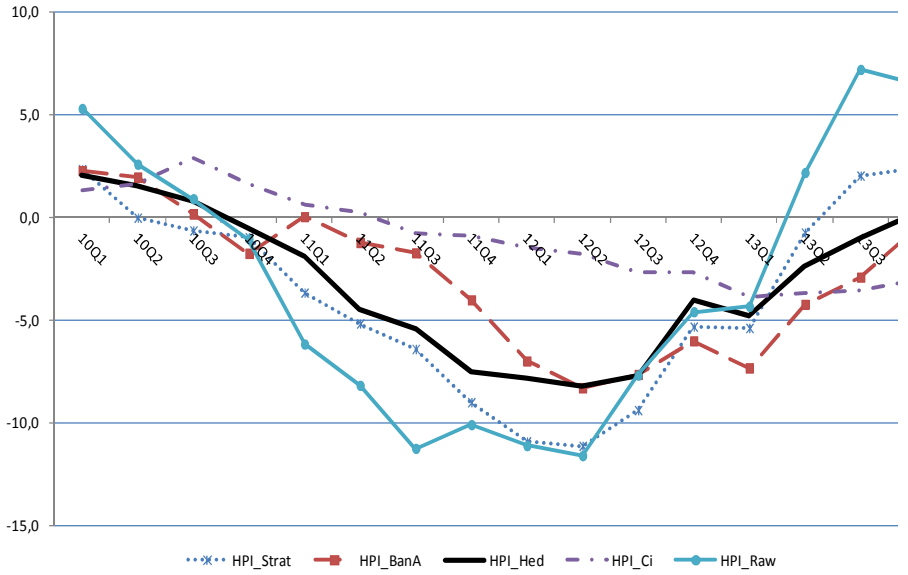
[18] Both the appraisals and transactions-based indexes signal the same turning point in 2Q2012. In addition, the correlation between the former and the latter indicators is stronger than the one found between the transactions-based index and the HPI that uses asking prices.

Chart 2: Index numbers and number of house sales (1Q2009-4Q2013)



	HPI_Strat	HPI_BanA	HPI_Hed	HPI_Ci
Mean	0.94	0.97	0.96	1.00
Stdev	0.07	0.09	0.06	0.02

Chart 3: Year-on-year changes (1Q2010 - 4Q2013)



	HPI_Strat	HPI_BanA	HPI_Hed	HPI_Ci
Mean	-3.9%	-3.0%	-3.2%	-1.0%
Stdev	4.67 p.p.	3.46 p.p.	3.53 p.p.	2.17 p.p.

asking-based price index and, on the one hand, the appraisals-based HPI and, on the other, the transactions-based HPI are 0.68 and 0.44, respectively. The correlation increases to 0.85 and 0.64 when appraisals and transactions year-on-year rates of change of quarter *Q* are compared with asking rates of change of *Q+1*. These results suggest the existence of a response lag in the asking-based HPI. In addition, when compared with the remaining HPIs, this indicator is less volatile and shows a different behaviour in the intensity of the price change in contraction/growth periods. In particular, it is possible to see that the index based on asking prices observes much lower price decreases in periods where the market is contracting[19]. Thirdly, the results reinforce the idea that the stratification approach does not fully take into account changes in the quality mix of transacted dwellings. This is suggested by the results shown in chart 3 where the HPI_Strat and HPI_Raw show higher volatility than the HPI_Hed price index. Finally, it should be underscored that house sales numbers are synchronized with the behaviour of appraisals- and transactions-based HPIs.

Conclusions and final remarks

Obtained results suggest that a HPI based on bank appraisals could be a reasonable proxy of a transactions-based price index. This is an interesting finding, especially if bank appraisals data are the only available information for the compilation of a HPI. However, it would be necessary to investigate more this issue in order to know if this outcome holds under all market conditions.

The work that was carried out reinforces the idea that asking-based indicators are less volatile, more resistant to price drops and, in addition, may lag behind transactions-based indexes when the housing market is depressed. This casts some doubts about the appropriateness of using these indicators for inflation monitoring purposes.

In general, the work reinforces the idea that it is possible to build good quality statistics and meet user's needs on real estate statistics using administrative data sources. Moreover, and in the specific case of Portugal, changing from appraisals to

[19] The behaviour of the asking-based HPI should come as no surprise, as this index is perhaps best representative of the beginning of the buying and selling process, where prices tend to be, especially in contraction periods, advertised at higher levels than actual transaction prices.

transactions would represent an improvement in relation to the methodological soundness, accuracy and reliability quality dimensions of already produced indicators. Indeed, this would not only stem from the use of transaction values (instead of bank appraisals) and from the quality of used fiscal data (e.g., it covers the complete universe of transactions) but also from the application of more appropriate methods for the treatment of the changing quality of transacted dwellings (i.e., the hedonic method).

References

[1] Confidencial Imobiliário (2006). *Índice Confidencial Imobiliário: Metodologia (Base 2005)*.

[2] Triplett. J. (2006). *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes: Special Application to Information Technology Products*. OECD. Paris.

[3] Rosen. S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation and Pure Competition. *Journal of Political Economy*.82. 34--55.

[4] Berndt. E. R. (1991). *The practice of econometrics: Classic and contemporary*. Addison-Wesley Publishing Company. Reading (MA).

[5] Eurostat (2011). *Residential Property Price Indices Handbook*. Available on http://epp.eurostat.ec.europa.eu.

[6] Ramsey, J.B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society*. B 31. 350--371.

[7] Kennedy. P. E. (1981). Estimation with Correctly Interpreted Dummy Variables in Semilogarithmitic Equations. *American Economic Review*. 71(4). 801.

[8] Kiel. K. And Zabel. J. (2008). Location. location. location: The 3L Approach to house price determination. *Journal of Housing Economics*. 17. 175--190.

[9] Goodman and Thibodeau (1995). Age-Related Heteroskedasticity in hedonic House Price Equations. *Journal of Housing Research*. 6(1). 25--42.