MEASURING THE EFFECT OF THE REAL ESTATE BUBBLE: A HOUSE PRICE INDEX FOR BILBAO

M.J. BÁRCENA, P. MENÉNDEZ, M.B. PALACIOS, AND F. TUSELL

1. Introduction

Since the accession of Spain to the EEC in 1986, and in particular during the first years of the present century, housing prices have experienced substantial increases. This phenomenon has been observed in a large part of the western world, but in Spain has been exacerbated by a number of circumstances: monetary stability, with low or even negative real interest rates, easy borrowing, high economic growth and fiscal allowances for home buyers. Since the first years of this century it became commonplace to refer to the *real estate bubble*.

The financial crisis started in 2007 brought an abrupt end to this state of affairs. Credit tightened, house sales slowed down to almost a halt, and a number of building companies went bankrupt or had to file for protection under the Spanish bankruptcy law. As the slowdown propagated to all sectors and unemployment began to rise, more and more home owners have been unable to meet their mortgage payments. All this has led to substantial downward pressure in the residential property market, and price cuts have ensued.

As prices began to drop, widely different figures were given on how much they had already dropped (and by how much they would drop in the sequel). Most interestingly, widely different figures were given on the extent of the price increase while the bubble lasted.

Part of the discrepancy can be traced to the fact that different sources sometimes speak of different markets (it is clear, for instance, that second residences in coastal areas have been hit harder than urban properties in or near large cities). But even when speaking of roughly the same market, figures are so widely disagreeing so as to raise the question of how they were obtained and what they really measure.

This paper is the first of a planned series which deal with the problem of measuring real estate property prices. We will use the housing market in Biscaye as a test case, proceeding from a mostly descriptive analysis

Date: November 17, 2011.

Key words and phrases. price indices, housing prices, semi-parametric models, geographically weighted regression, GWR.

Partial support from grants ECO2008-05622 (MCyT) and IT-347-10 (Basque Government) is gratefully acknowledged.

confined to the city of Bilbao in this report towards more theoretical and methodological issues in those to follow.

The remaining of this paper is organized as follows: Section 2 gives an overview of available sources of housing prices in Spain and the institutional framework. Section 3 introduces a semi-parametric and geographically weighted hedonic model which, among other things, produces a price index estimate. Section 4 compares our price index with related indices from alternative sources and draws some conclusions.

2. Housing prices and indices.

2.1. **Methodological problems.** The construction of price and quantity indices is one of the main tasks undertaken by all statistical offices. As such, it has been carefully scrutinized and the properties of different methods to construct indices are now well understood: see for instance Vogt and Barta (1996).

Laspeyres indices are one of the most popular choices. Given a basket made of quantities q_{i0} which at the base period $s = 0$ had prices p_{i0}, $i = 1, \ldots, I$, the index for period $s = t$ with base at period $s = 0$ is computed as

$$I_0(t) = \frac{\sum_{i=1}^{I} p_{it} q_{i0}}{\sum_{i=1}^{I} p_{i0} q_{i0}} \tag{1}$$

This requires that items $i = 1, \ldots, I$ are well standardized and observable over time. Since some items may disappear and be replaced by similar ones, formula (1) is only usable over a limited period: chained Laspeyres indices are used instead.

Houses are traded at relatively infrequent times and can hardly be standardized: two equally built and furnished houses may command widely different prices in the market on account of their different location or even orientation. Clearly, the computation of an index such as (1) is unfeasible, because no given set of houses are traded at regular intervals, and no exact replicates are available.

Both problems —lack of homogeneity of houses and irregular observation times— have been addressed in a variety of ways, which include the estimation of hedonic models, construction of "cells" of relatively similar houses and repeated sales methods (cf. Rodríguez López (2007)). Some of the statistical sources mentioned in Section 2.3 attempt to construct categories of homogeneous houses; the modelling approaches are briefly described next.

Hedonic models. Hedonic models go back to at least the celebrated paper Court (1939). The paper Rosen (1974) is usually credited as laying down a sound theoretical basis. A useful collection of papers dealing with hedonic models in housing markets is Baranzini et al. (2008).

Hedonic models attempt to measure the contribution of individual characteristics to the total value of the house; each house is therefore regarded as a "basket" of characteristics like size, quality of the construction, age and the quality of the surroundings (e.g., clean air or noise).

Location is typically included in the form of proxies which measure distance to transportation networks, municipal services, recreational areas, etc.

Since this proxies are not always observable and rarely capture fully all factors affecting the desirability of a given location, spatial effects are likely to remain; this can be introduced through the error structure, e.g.. Dubin (1988), as additional effects in the regression (typically in the form of smooth surfaces over space, non-parametrically estimated, as in Hastie and Tibshirani (1991)) or by allowing the coefficients of the hedonic model to change in space: this approach, named geographically weighted regression (cf. Fotheringham et al. (2002)), will be followed in Section 3 (see also Kestens et al. (2005)).

Hedonic models have also been used to measure the effect of time on selling prices of homes, which is of particular interest here. Time is not necessarily viewed as a cause, but as a proxy of a variety of factors which change over time and cannot be explicitly accounted for. The specification typically takes log price as the response variable in a regression model; the estimated coefficients of time dummies are used to measure the cumulative percentage of change in constant quality house prices up to and including the associated time period.

One advantage of hedonic models is that all home sales observed in a time period can be used in the estimation. By contrast, repeat sales house prices can be used; in that case only houses which have been sold at least twice within the observed period can be used in the estimation. Quality adjustment is simple and automatic in repeated sales models[1], as we consider differences in price of the same houses; the downside is that the sample size is reduced to houses with more than one trade.

2.2. **Institutional framework.** The market of new houses in Spain is split into the *free* and *protected* segments. House prices in the free segment are not subject to any restrictions whatsoever. The protected segment *(vivienda protegida)* is subject to ceilings in price, and is meant to guarantee a fraction of affordable houses for people who would otherwise be unable to buy one. A percentage of all new house starts must be in the protected segment[2]. After the first transmission, protected houses are traded much as free houses: the authorities have a preferential buying right of any protected house offered in the market, but only exceptionally has this right been exercised.

A feature of the real estate market is its opacity: prices really paid in real state transactions are rarely known with any degree of confidence. There are incentives to declare different (in general, lower) prices than those really paid, as taxes on the transactions are assessed on a per value basis[3].

[1]Assuming house characteristics remain constant between consecutive sales dates.

[2]In the Basque Country, as of this writing, 20%.

[3]New houses are subject to the Value Added Tax (IVA), at a rate of 7% for most of the period analyzed, then raised to 8% of their value. In mid 2011 the IVA rate for new houses was lowered to 4%, in an attempt to boost the languishing housing market. Second hand houses pay a tax on transmissions of 6%. On the other hand, the seller must pay income tax on any increase of value of the sold property, although allowance is made for

Since real transaction prices are not readily observed, approximations are used. Declared prices, as mentioned, are usually different (and below) real prices. A different source of approximate prices are the value assessments made by specialized agents. These value assessments are collected nation-wide[4] and are one of the inputs for the Survey of Housing Prices[5] described below. For the free segment, these assessments of market value are usually required by banks from prospective buyers who wish to finance their buy with a mortgage[6].

The use of market value assessments, however professionally made, raises several objections. First, there has been a tendency to overstate the value of properties in order to disguise loans which otherwise would appear insufficiently backed by the value of the mortgaged property. Second, it is not usually the case that a market value assessment is done on a house which is not mortgaged: this may lead to selection biases, as only part of the transactions are reflected in value assessments (cf. García Montalvo (2007)).

All these factors taken together explain that statistics on housing prices are sometimes difficult to reconcile, as they are based either on declared or assessed prices. We briefly describe some statistical sources of information next.

2.3. **Statistical sources.** There are several sources of statistical information on house prices in Spain. The Ministerio de la Vivienda, or Ministry of Housing (MV), and the Instituto Nacional de Estadística, or National Statistical Institute (INE), both provide information on housing market conditions and prices; the INE publishes the Índice de Precios de la Vivienda, or Housing Price Index (IPV), whose methodology is harmonized with that used in other countries of the European Union (EU). Other sources of information include real estate agencies and firms specializing in value assessment of real estate.

Housing Prices Survey. The Estadística de Precios de la Vivienda, or Housing Prices Survey (EPV), offers nation-wide statistical information on house prices and quantities provided by the MV[7].

The MV addresses both the protected and free segments. The source data for the protected segment are the maximum legal prices. For the free segment, market value assessments are used. The assessed market value is defined as

---

the effect of inflation. Both factors provide an incentive to understate the value of the transaction. Finally, there is a local tax on the increase of value of land, although this last is levied with no reference to the declared value of the transaction, and therefore provides no further incentive to distort declared prices.

[4]By ATASA, Asociación Profesional de Sociedades de Valoración.

[5]Estadística de Precios de la Vivienda, Ministerio de la Vivienda.

[6]The maximum amount of money to be lend is usually 80% of the assessed value of the property, although in the midst of the bubble 100% was sometimes granted.

[7]The information to follow is summarized from an undated report, Ministerio de la Vivienda (retrieved 10-June-2010).

> "...the price at which a property could exchange hands at the date of the assessment, exclusive of taxes and commissions."

Therefore, these are *not* prices at which transactions have taken place.

From the assessed market values, the MV produces indices using a bottom-up approach. Prices of homes of a given type (new or second hand, protected or free) are aggregated for each price stratum[8] and area and divided by the total built surface[9] to produce an average price in €/m^2. These are further weighted and aggregated to produce indices for larger areas.

The price indices derived from the EPV are published quarterly.

Assessment firms and real state agencies. There are at least two price indices claiming fairly general coverage.

The Índice de Mercados Inmobiliarios Españoles, or Spanish Real Estate Market Index (IMIE) is an initiative of a private company[10] based on information from over 200.000 value assessments, collected by qualified staff with daily contact with the market. It is published monthly, and is a chained Laspeyres index, with base year 2001.

The spatial resolution tends to adapt to market segments rather than administrative boundaries: large cities, metropolitan areas, Mediterranean coast, Balearic and Canary Islands and Rest of municipalities. These market segments are seen as the core of the real state market, leaving into relative oblivion rural areas.

Aside from the use of assessed values rather than transaction prices, sample selection is to be feared, as assessment is more likely for properties to be mortgaged, and this may bias the sample towards some market and price segments.

Another private firm which publishes an index of the Spanish real state market is Sociedad de Tasación[11]. It produces an index of nominal prices per square meter, base year 1985. It is published twice per year and gives information for the whole of Spain as well as Madrid, Barcelona, Autonomous Communities and some large cities.

Several web sites[12] offer price information with different levels of processing and aggregation .

Housing Prices Index. The last addition to the sources of statistical information on housing prices is the IPV, an statistical operation whose aim

[8]Seven strata are defined, with a top price of 1.050.000€; thus, the most expensive houses are excluded.

[9]Built surface *(superficie construida)* includes closed balconies, and structural vertical elements, as opposed to usable surface *(superficie útil)*, which excludes those elements as well as 50% of balconies under roof; for a more precise description, see Ministerio de la Vivienda (retrieved 10-June-2010), p. 7. A useful rule of thumb is that usable surface is usually about 15% less than built surface.

[10]TINSA Consultancy and Environment, `http://www.tinsa.es`.

[11]It specializes in value assessments of all types, but particularly of real state assets. Present in the market since 1982, see `http://web.st-tasacion.es`.

[12]See for instance `http://casas.facilisimo.com/preciometro` and `http://www.idealista.com/pagina/informes-precio-vivienda`.

and methodology is closest to our work. The IPV is a housing price index published quarterly by the INE with base year 2007. Its primary objective is to measure the evolution over time of house prices in both the new and second-hand free segments. It thus excludes prices in the protected segment. The IPV aims to serve as a comparable statistical source within the scope of the European Union harmonized statistics.

The document INE (2009) contains a fairly detailed discussion of the decisions made on primary sources to use. The basic information source retained is the database provided by the General Council of Spanish Notaries (Consejo General del Notariado). Virtually all house transactions are made with the intervention of a notary public. The transaction price along with the date, characteristics of the house (flat or single family house, built surface, new or second hand, parking space if available, etc.) is transmited to the INE each month. Prices are therefore prices declared by both parties, buyer and seller, when signing the transfer of property. This need not be the real amount paid since, as explained in the previous section, there are incentives to understate (and, in some cases, to overstate) prices; however, it has been deemed the best alternative.

The IPV fully recognizes that quality of the houses transacted has to be taken into account. The approach followed involves the use of an hedonic model involving the variables mentioned above and additional information such as size of the municipality, type of environment, and the degree to which the location of the house is touristic. The dependent variable is the log price per square meter.

All house characteristics entering the hedonic model are qualitative[13]. The main effects they define along with selected interactions involve 157 parameters that have to be estimated each quarter. Each combination of explanatory variables defines one cell[14]. The model yields average price per cell which is then weighted into a general price index and price indices for different aggregations of cells[15].

If we denote by $\boldsymbol{x_c}$ the vector of values of the explanatory variables (main effects and interactions) defining cell c, and $\ell_{i,c,q}$ the log price of house i in cell c at time (quarter) q, the model is:

$$\ell_{i,c,q} = \boldsymbol{x_c}\boldsymbol{\beta^q} + \epsilon_{i,c,q} \tag{2}$$

Stacking (2) for all observations corresponding to one quarter, we obtain:

$$\boldsymbol{L_q} = \boldsymbol{X_q}\boldsymbol{\beta^q} + \boldsymbol{\epsilon_q} \tag{3}$$

which is re-estimated every quarter q, yielding a time-varying vector of parameters $\boldsymbol{\beta^q}$.

[13]The built surface is categorized.

[14]For the years 2007 and 2008, nearly 52.000 such cells were used.

[15]The weights are given by the ratio between the value of houses transacted in the cell and the total value of houses transacted. Hence, selection biases may also creep in.

INE (2009) gives further details, concerning imputation of missing data, correction of heteroskedasticity and bias removal, so prices for each cell (rather than log prices) are approximately unbiased.

Other sources of information. The Basque Government publishes quarterly a bulletin, Departamento de Vivienda Obras Públicas y Transportes (2011), with a wealth of information (and commentaries) regarding trends not only in prices, but also on availability of housing, house starts, land prices, developable land, rents, etc.

Some web sites which provide real estate advertising and services also publish average prices and price indices, as a by-product obtained from the information they have access to. Among them, `www.idealista.com`, whose raw data we use in the models estimated in Section 3.

3. Modelling price changes

3.1. **Data.** We have used data obtained from `idealista.com`, a leading web site in the area of real state. Data consists of house offers in the city of Bilbao, from 2004-11-17 to 2011-05-04; in all, 7524, of which 6696 contained full street addresses that we have been able to geocode[16]. Data consists of a description of the property offered for sale, including selling price, location (often down to the street number, in other cases only street or district), square footage, floor, age of the building, number of bedrooms and bathrooms, parking space if available, availability of central or individual hot water and heating, elevator, etc. We emphasize that these are offered or selling prices, not transaction prices. While in principle transaction prices would be desirable, selling prices appear to be good proxies and have been used in a number of studies, including Kryvobokov and Wilhelmsson (2007), Henneberry (1998) and Pace et al. (2000).

Figure 1 shows a polygon outline with a map of Bilbao[17] as background, for reference. To avoid distraction, all subsequent plots of results will be made over the outline. Figure 2 shows the approximate location of districts referred to later.

We have fitted several models, whose relative merits are discussed next. It is remarkable that different specifications, using location information with various degrees of sophistication, still yield fairly similar estimates of the price index.

3.2. **Model 1.** Our first model is a standard hedonic model in which the response is $\log(\text{Price}/\text{m}^2)$. A categorical variable codes the district where each offered house is located; there are 13 such districts in Bilbao. Other (fixed) effects introduced are type of dwelling (flat, duplex, etc.), type of heating (or unavailability of heating, as the case may be), number of parking

[16]But note that a smaller number of observations has been used to estimate the models below, as observations with missing variables may be discarded.

[17]From Google Maps, `http://maps.google.com`.

FIGURE 1. Outline of the city of Bilbao

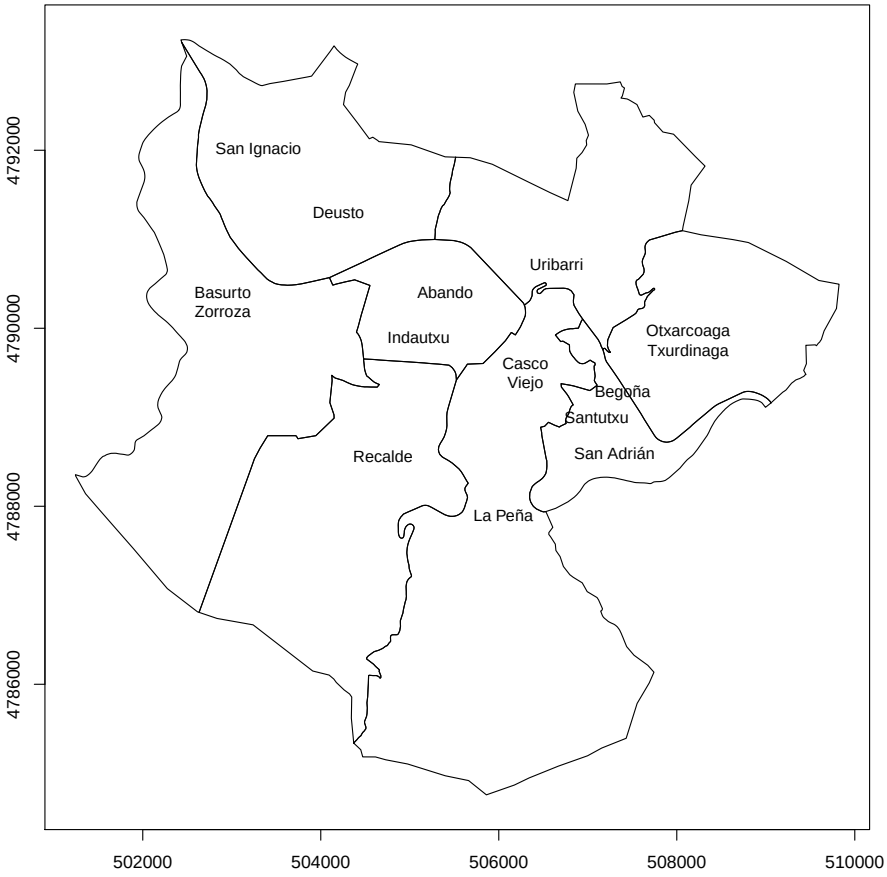
places, number of bathrooms, of bedrooms, whether there is elevator, the monthly cost of shared services and age of the building, coded in 7 categories.

The effect of time is modeled non-parametrically, as a cubic spline with 12 degrees of freedom. Thus, Model 1 can be written as:

$$\log(\text{Price}/\text{m}^2) = \sum_{i=1}^{p} \beta_i x_i + s(t) + \epsilon \tag{4}$$

where the x_i $(i = 1, \ldots, n)$ are the observed regressors, $s(t)$ is a cubic spline whose suitably normalized value provides our estimation of the price index and ϵ is a random disturbance. Values of the estimated parameters can be seen in Table 1.

FIGURE 2. Districts of Bilbao. The axis units are UTM coordinates within the sheet 30N, in meters.



Models such as (4) can be readily fitted with off-the-shelf software. We have used the `gam` function in the `R` package `mgcv` (see Wood (2001) and Wood (2004)). Results are shown in Table 1. Since the response is measured in the log scale, $\exp(s(t))$ enters as a factor in Price/m^2 and our estimation for the index with base 100 at time t_0 is obtained as

$$I(t) = 100 \times \frac{\exp(s(t))}{\exp(s(t_0))};$$

the profile of $I(t)$ along with a 95% confidence interval are shown in Figure 3. We see that offered prices reached their maximum around the end of 2007, accumulating an increase of almost 25% since the beginning of 2005, then dropped until the first quarter of 2009 at which time they appear to have stabilized slightly above the early 2005 levels, only to later resume their drop.

TABLE 1. Estimation results for Model 1.

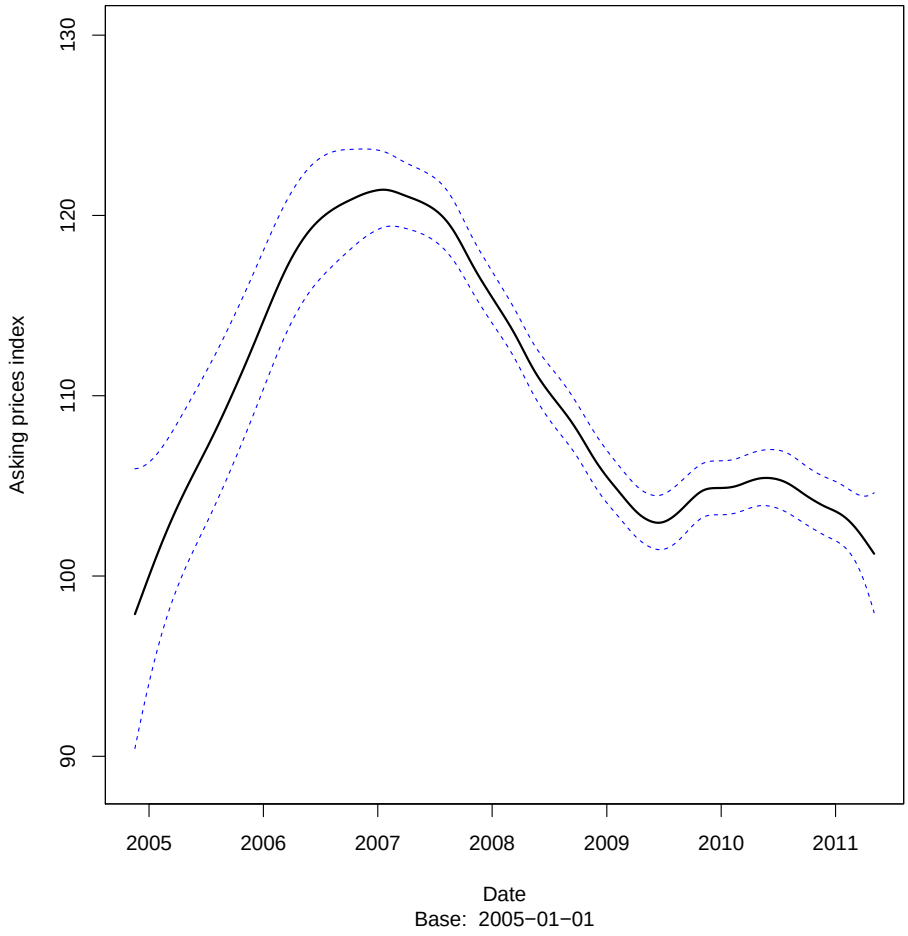
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.6721	0.0503	192.34	0.0000
ExteriorTRUE	0.0120	0.0079	1.52	0.1290
Bedrooms	-0.0168	0.0040	-4.19	0.0000
Byears	0.1055	0.0064	16.50	0.0000
Parking places	0.0737	0.0068	10.82	0.0000
log(Square footage)	-0.3384	0.0134	-25.18	0.0000
Community fees	0.0007	0.0001	7.32	0.0000
ElevatorTRUE	0.1859	0.0065	28.68	0.0000
Type of heating:				
colective	-0.1058	0.0132	-8.03	0.0000
unstated	0.0398	0.0073	5.48	0.0000
individual	-0.1413	0.0081	-17.48	0.0000
unavailable	-0.0012	0.0470	-0.03	0.9795
Type of house:				
duplex	-0.0505	0.0272	-1.86	0.0636
studio	-0.1764	0.0369	-4.78	0.0000
flat	-0.0411	0.0109	-3.77	0.0002
single house	-0.1565	0.0680	-2.30	0.0214
house in a block	-0.0141	0.0436	-0.32	0.7460
paired house	-0.0632	0.0564	-1.12	0.2620
District:				
basurto - zorroza	-0.0954	0.0091	-10.52	0.0000
begoña - santutxu	-0.0466	0.0100	-4.67	0.0000
casco viejo	0.0794	0.0124	6.41	0.0000
deusto	-0.0133	0.0114	-1.17	0.2439
ibaiondo	-0.1394	0.0088	-15.93	0.0000
indautxu	0.2196	0.0105	20.85	0.0000
otxarkoaga - txurdinaga	-0.1561	0.0156	-10.01	0.0000
rekalde	-0.1352	0.0153	-8.86	0.0000
san adrián - la peña	-0.0462	0.0085	-5.41	0.0000
san ignacio	-0.1573	0.0138	-11.37	0.0000
uribarri	-0.0724	0.0154	-4.70	0.0000
Age of building:				
$<$ 5 years	0.0960	0.0120	8.02	0.0000
5-10 years	0.0292	0.0147	1.99	0.0469
10-20 years	-0.0112	0.0077	-1.46	0.1452
20-30 years	0.0685	0.0132	5.19	0.0000
+ 30 years	0.0220	0.0078	2.82	0.0048
Time trend:				
s(x, 12)	-0.0001	0.0000	-20.83	0.0000

Models such as (4) can be readily fitted with off-the-shelf software. We have used the `gam` function in the `R` package `mgcv` (see Wood (2001) and Wood (2004)). Results are shown in Table 1. Since the response is measured in the log scale, $\exp(s(t))$ enters as a factor in Price/m^2 and our estimation for the index with base 100 at time t_0 is obtained as

$$I(t) = 100 \times \frac{\exp(s(t))}{\exp(s(t_0))};$$

the profile of $I(t)$ along with a 95% confidence interval are shown in Figure 3. We see that offered prices reached their maximum around the end of 2007, accumulating an increase of almost 25% since the beginning of 2005, then dropped until the first quarter of 2009 at which time they appear to have stabilized slightly above the early 2005 levels, only to later resume their drop.

FIGURE 3. Housing price index for Bilbao: non-parametric estimate from Model 1. Base 100 January, 2005. Dotted lines give 95% point wise confidence intervals.



Let P_i^* be the deflated price per square meter of house i. Next step is to fit a geographically weighted regression,

$$\log(P_i^*) = \sum_{j=1}^{p} \beta_{i,j} x_{ij} + \epsilon_i \tag{5}$$

where x_{ij} is the value of the j-th regressor for house i. Coefficients $\beta_{i,j}$ are estimated for each regressor j at each house location i[18].

The estimation is done by least squares, with observations weighted less as their distance to house i increases. Weighting is usually done with a two dimensional kernel, here an isotropic gaussian kernel whose bandwidth is selected by cross-validation. With the sample used, the bandwidth selected[19] (= standard deviation of the gaussian kernel) has been 350.8, which implies that observations 350.8 meters away are given weights about 6% of those right at the space point at which we estimate the coefficients ($6\% \approx 0.0585$, density of an isotropic bivariate normal density one standard deviation away from the mean). The cross-validated bandwidth is fairly small: in another setting, Can and Megbolugbe (1997) found that prices at one location are affected by prices within about a two mile radius.

Results of Model 2 are not presented here, as it is only a logical step towards Model 3, described next.

3.3. **Model 3.** Model 2 in the preceding section is meant to capture the effect on prices of house characteristics in spatially dependent form at a given point in time. This is not our primary goal: what we want is to adjust prices for house quality and location, for the purpose of obtaining a constant quality price index. What we would like instead is a model such as

$$\log(P_{it}) = \sum_{j=1}^{p} \beta_{i,j} x_{ij} + s(t) + \epsilon_{it} \tag{6}$$

in which the linear coefficients $\beta_{i,j}$ are as before hedonic coefficients spatially varying, P_{it} are current prices per square meter and $s(t)$ is a smooth function capturing the evolution of prices. In a sense, (6) is a merger of Model 1 and Model 2, in that we have both a non-parametric estimate of the price level and spatially varying hedonic terms.

We are not aware of software enabling the fitting of models such as (6) off-the-shelf; however, given routines to estimate Models 1 and 2 it is fairly straightforward to implement a back-fitting estimation routine (cf. Hastie

[18]Although for simplicity we are assuming that we estimate parameters only at locations where we observe a house, this need not be the case. We can estimate parameters at any point in space, irrespective of whether or not an observation is present there. In particular, we might estimate parameters over a grid of points to construct maps by smoothing and interpolation of the obtained values.

[19]As the geocoding is made to UTM coordinates in meters, the bandwidth is also in meters.

and Tibshirani (1991), § 4.4) for Model 3. The procedure can be sketched as follows:

Step 0. Take as initial estimate $s^{(0)}(t)$ of $s(t)$ in (6) the smooth function estimated from (4) (or any other available approximation).

Step 1. At iteration k, let $\beta_{ij}^{(k)}$ be the estimates of the $\beta_{i,j}$ using geographically weighted regression to fit

$$\log(P_{it}) - s^{(k)}(t) = \sum_{j=1}^{p} \beta_{i,j}^{(k)} x_{ij} + \epsilon_{it}^{(k)} \tag{7}$$

Step 2. At iteration k, obtain $s^{(k+1)}(t)$ by smoothing over time the partial residuals $\log(P_{it}) - \sum_{j=1}^{p} \beta_{i,j}^{(k)} x_{ij}$.

Step 3. For a pre-set tolerance η, if $\max |s^{(k)}(t) - s^{(k+1)}(t)| < \eta$, return as estimates the $\beta_{i,j}^{(k+1)}$ and $s^{(k+1)}(t)$ computed in the final iteration, otherwise return to Step 1.

Essentially, the back-fitting algorithm iterates Step 1 and 2, each time estimating the parametric or non-parametric part of the fit given the best current approximation of the other.

As it happens, convergence is attained in just three iterations; the starting (from (4), Model 1) and final (from (6), Model 3) estimates of the price index are shown in Figure 4. They are not very different from each other; the back-fitted index closely follows its counterpart from Model 1 until mid 2006, and stays above ever since, albeit by a small amount. For the purpose of comparison, we plot also the IPV from INE (refer to section 2.3) for the whole of the Basque Country[20]. (Since the IPV is computed with base 2007 = 100, we have aligned their first published figure with the half sum of our two indices.)

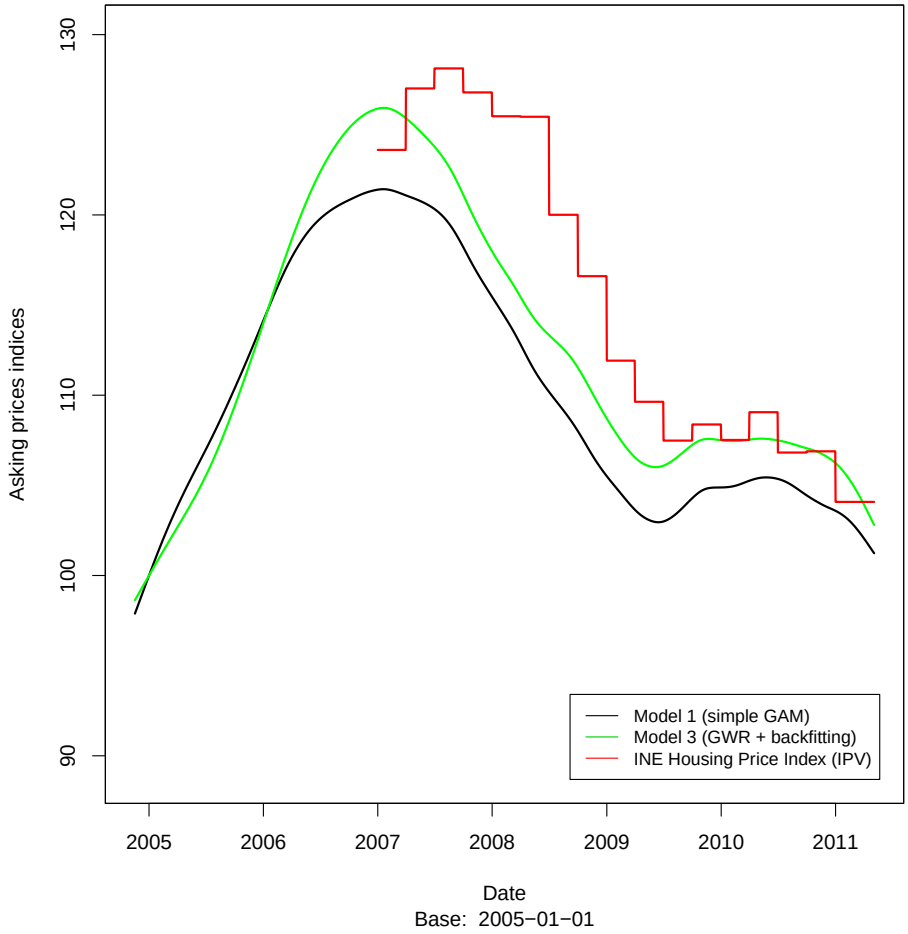
Although the IPV is computed for a larger area and with different methodology, its profile is remarkably similar to our two indices. In particular, all of them nicely reflect a short-lived upturn in early 2010, likely due to the announced end of fiscal rebates for house buyers, which prompted a small surge of activity in the housing market.

The similar profiles of the IPV and asking prices indices are also significant in that they help to allay fears that they could measure entirely different things. One might hypothesize that asking prices are relatively sticky, and softening market conditions would surface rather in transaction prices (which are input to the IPV). This does not seem to be the case.

Model 3 produces a wealth of geographically disaggregated information. The parametric part (i.e., the right hand side of equation (7) evaluated with the estimated $\beta_{i,j}^{(k)}$), suitably scaled, gives estimates $\log(P_{it}/m^2) - s(t)$, i.e. deflated log prices per square meter. Aligning values to the base period and

[20]Obtained from their web page, `http://www.ine.es`, visited July, 14, 2011.

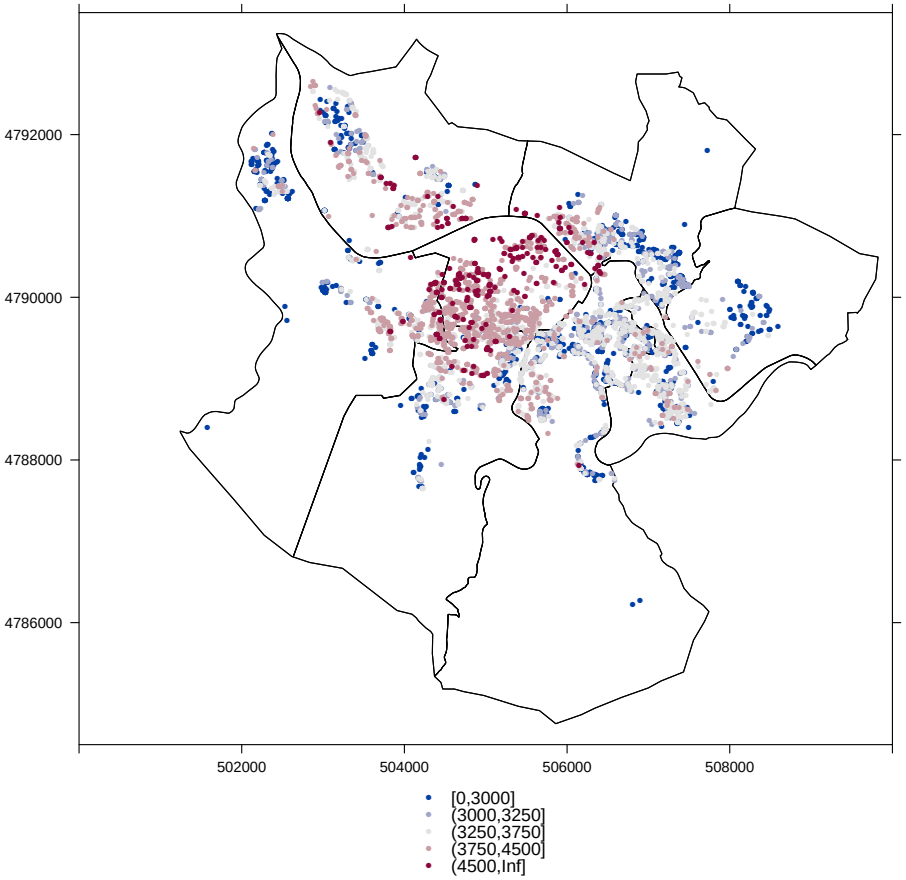
FIGURE 4. Asking prices indices. The index from INE (base 2007) has been aligned with the average of our two indices at 2007-02-15, mid point of the first quarter for which it was published.



transforming back to the original scale (€/m^2), we get the values mapped in Figure 5. (We have made no attempt to correct biases due to the non-linear transformation.) The pattern in Figure 5 is clear; houses in the central part of the city command the highest prices, while peripheral districts of San Ignacio, Txurdinaga or Basurto exhibit lower prices.

It is also interesting to check the fit, which now is local. Since we compute a different regression for each location, we have local R^2 coefficients that we have plotted in Figure 6. The fit of deflated log prices per square meter with the available regressors is modest, with the median R^2 equal to 0.4497.

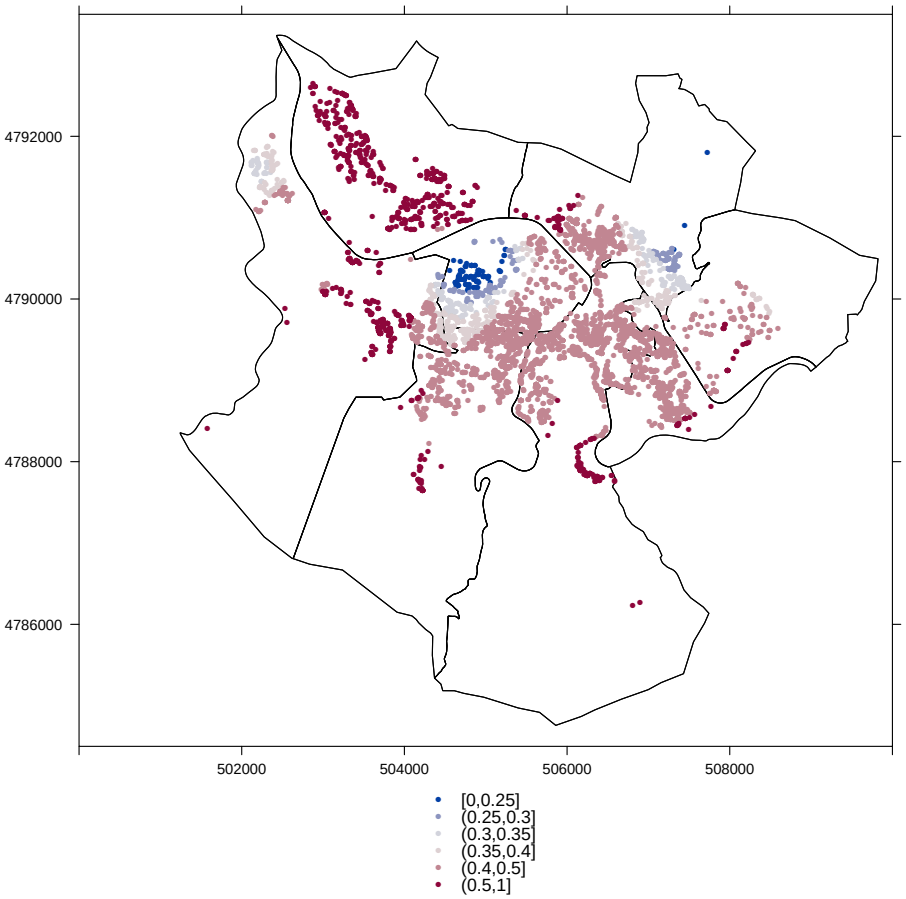
FIGURE 5. Fitted values in € per square meter at the base period (January, 2005).



There are localized areas of worse and better-than-average fit for which we have no obvious explanation.

In principle, local values for any of the $\beta_{i,j}^{()}$ coefficients in equation (6) could be spatially represented. We could then have an estimate of how much e.g. an additional bedroom, existence of elevator or a garage, enhances the price of a house at each chosen location. This may be misleading: as pointed out in Wheeler and Tiefelsdorf (2005), local multicollinearity is usually a problem. Some remedies have been proposed (e.g., Wheeler (2006), Vidaurre et al. (2011)) which we have not attempted to implement, as for the purpose of price index construction we only require a good estimate of $\sum_{j=1}^{p} \beta_{i,j}^{(k)} x_{ij}$ in (7), and are not concerned with apportioning that sum into the contributions of each predictor.

FIGURE 6. Local (partial) R^2 values for the regression $\log(P_{it}) - s^{(k)}(t) = \sum_{j=1}^{p} \beta_{i,j}^{(k)} x_{ij} + \epsilon_{it}^{(k)}$.



4. DISCUSSION

Our goal was to assess the feasibility of building a house price index with publicly available information on proxy variables of the true magnitudes to estimate. This is quite in keeping with the modern trend to exploit administrative records or automatically collected information to replace, to the extent possible, costly surveys.

The limited evidence shown in the precedent section is encouraging: with off-the-shelf software and limited human resources (basically spent in data cleaning and geocoding) we have been able to compute an index which

captures remarkably well the patterns in the IPV[21]. The method is straightforward, simple to implement, understand and track.

However, ours is but a possible approach to a much researched problem: for a collection of papers from a spatio-temporal vantage point see Biggeri and Guido Ferrari (2010). Specifically related to housing prices and close to our approach are, among many others, Borst (2008), Geniaux and Napoléone (2008), Clapp (2004), Meese and Wallace (1997) Bourassa et al. (2006), McMillen (2011) and McMillen and Redfearn (2007). Comments on some of their respective approaches follow, to highlight similarities or differences with the approach followed in Model 3; Table 2 gives a summary. We can make no attempt to even list all contributions in this area, which has produced a copious literature.

McMillen (2011) goal is to examine the effect over time on house prices of the distance to "employment sub-centers" in the city of Chicago. He uses a repeated sales estimator and a smooth function to capture the effect of time. His model takes the form

$$Y_{i,t} - Y_{i,s} = g(t) - g(s) + u_{i,t} - u_{i,s}, \tag{8}$$

where $Y_{i,t}$ is the (log) price of house i at time t, u_{it} the corresponding random term and $g(t)$ is a smooth function of time given by

$$g(t) = \alpha_0 + \alpha_1 z + \alpha_2 z^2 + \sum_q \left(\lambda_q \sin(qz) + \gamma_q \cos(qz)\right), \tag{9}$$

with $z = 2\pi t/\max(t)$. As compared with our model, he does not need to fit a hedonic model, as the effect of the quality of house i cancels in the difference $Y_{i,t} - Y_{i,s}$ (assuming the characteristics of the house i remained constant from time s to time t); the downside is that only houses with repeated sales in the sample can be used (with a possible selection effect, as poor quality houses might be more likely to enter the market repeatedly). To capture the effect of time, a second order polynomial plus a combination of periodic functions is used. The profile of $g(t)$ suitable normalized would give an estimate of a price index similar to ours, although for a given number of degrees of freedom we regard the penalized spline approach we have used as more flexible.

Geniaux and Napoléone (2008) are not concerned with the effect of time, but rather spatial effects and in particular the distance effect to urban center. In their paper they examine the effect of distance to Avignon on prices of rural properties in the south of France. They consider a variety of models, like

$$Y_i = \beta X_i + Z_i * s(u_i, v_i) + s(u_i, v_i) + \epsilon_i \tag{10}$$

[21]It is important not to be over-optimistic when interpreting Figure 4. One might be tempted to claim that our index appears to anticipate movements in the IPV. This would not be a fair claim, as the IPV is computed with only past and present information, while our index uses also *future* information in the smooth.

where Y_i is again the log price of house i, X_i captures characteristics that have a fixed (across space) effect, and Z_i is a variable whose effect is space-dependent; it is multiplied by a "smooth" function of geographical coordinates to capture such dependency. They go on considering alternatives such as generalized additive models (GAM), geographically weighted regression (GWR) and mixed GWR (MGWR). They conclude that

> "GAM fits better than MGWR, is even more flexible in articulating stationary and non stationary coefficients, works well with a big sample and makes investigating non linearity easy." (p. 125)

While they are not concerned with the problem of modelling the effect of time (which enters some of their models only in the form of year dummies), their models are close to ours; in their Section 5.4.3 they propose an estimation method for the MGWR model which is also close to the back-fitting algorithm sketched above for Model 3.

Close to this approach is Clapp (2004), which proposes the model,

$$Y_{it} = \beta X_i + s(u_i, v_i, t) + \epsilon_{it} \tag{11}$$

where, as before, Y_{it} is the log of the transaction price, X_i is a vector of attributes of the i-th house in the sample, β the vector of implicit hedonic prices, and $s(.)$ a function of latitude, longitude *and* time. They consider fixed betas in the hedonic part and a smooth function in time and space (estimated by local polynomial regression, using a product kernel) to account for trends in time and space. The estimated $s(u_i, v_i, t)$ as a function of time for given u_i, v_i gives a *local* price index. The estimation of (11) is made in a two-step algorithm, using theory developed by Robinson (1988) and Stock (1989).

Meese and Wallace (1997) compare several methods, providing a wealth of insight on their respective merits. They propose yet another model,

$$Y_{i(t)} - \overline{Y}_t = G(x_{i(t)}) + u_{i(t)}; \tag{12}$$

Equation (12) is estimated for each quarter; $Y_{i(t)}$ is the log price of the i house traded at quarter t, and $\overline{Y}_t$ is the average log prices of all houses traded in that same quarter. $G(x_{i(t)})$ is the hedonic part, estimated by locally weighted regression (LWR); weighting is done not in terms of geographical distance, but rather distance in the attribute space to the attributes median. The implicit prices of the attributes are then used to correct $\overline{Y}_t$ with the valuation of the attributes of the current or first quarter. This gives Laspeyres and Paasche indices whose geometric mean produces a final Fisher's ideal index.

McMillen and Redfearn (2007) contains an enlightening discussion on kernel regression, conditionally parametric regression (CPAR) and locally weighted regression (LWR), which they show encompasses the previous methods and GWR as particular cases. An interesting approach that uses *both* sales prices and appraisal prices is described in Bourassa et al. (2006),

TABLE 2. Summary and comparison of some models and indices discussed

MODEL OR INDEX	QUALITY ADJUSTMENT	SPATIAL EFFECTS	TIME EFFECTS
Geniaux and Napoleone(2008): $Y_i = \beta X_i + Z_i s(u_i, v_i) + s(u_i, v_i) + \epsilon_i$	hedonic, fixed β	smooth $s(u_i, v_i)$	None
McMillen(2011): $Y_{i,t} - Y_{i,s} = g(t) - g(s) + \epsilon_{i,t} - \epsilon_{i,s}$	repeated sales	not explicitly	$g(t)$
Clapp(2004): $Y_{it} = \beta X_i + s(u_i, v_i, t) + \epsilon_{it}$	hedonic, fixed β	joint smooth $s(u_i, v_i, t)$	
Meese and Wallace(1997): $Y_{i(t)} - \overline{Y}_t = G(x_{i(t)}) + \epsilon_{i(t)}$	hedonic	LWR† in attributes	estimated for each t
Housing Prices Survey (IPV): $Y_{i,c,t} = X_c \beta_t + e_{i,c,t}$	hedonic, variable β	not explicitly	estimated for each t
Model 3 (Section 3.4 above): $Y_{it} = \beta_i X_{it} + s(t) + \epsilon_{it}$	hedonic, variable β	GWR‡	smooth $s(t)$

† LWR = Locally weighted regression.
‡ GWR = Geographically weighted regression.

which also contains a comparison among several versions of hedonic models, repeated sales methods and their SPAR (Sale Price Appraisal Ratio) method.

Our Model 3 produces an index for the whole area sampled, rather than indices for each particular location —like e.g. Clapp (2004). Where other models, such as Geniaux and Napoléone (2008), use smooth nonparametric functions in geographical coordinates to account for spatial effects, our model accounts for spatial variation through GWR of hedonic coefficients, the $s(t)$ term smoothing along the time dimension, much like McMillen (2011). Our model is geared towards production of an index for a region, rather than the discovery of sub-regions with different price trends.

Aside from the papers mentioned, there is a huge literature on spatio-temporal processes, potentially useful for our purposes. An up-to-date comprehensive overview is given in Cressie and Wikle (2011); see also Banerjee et al. (2003) and Gelfand et al. (2010). Specific applications to housing prices already exist that make use of this approach, e.g. Gelfand et al. (2004). For the purpose of computing price indices, we find hierarchical models (whether using a "classic" parameter estimation or a fully Bayesian approach) a highly attractive alternative, for the reasons outlined in Cressie and Wikle (2011), § 2.1; they afford a nice separation between the "process model",

describing how relevant magnitudes (or state variables) evolve, and the "data model", describing how observations are generated. Since our concern is to estimate (unobservable) latent variables such as the price level, this separation is meaningful; we may confine the latent variables to the "process model" and build the data model on top of it.

Specialized software exists that can be used to obtain posterior distributions of the magnitudes of interest through MCMC, if one is willing to follow a Bayesian approach (e.g. Brian J. Smith (2008)) although this is computationally demanding for the problem sizes we envision. Alternative methods like integrated nested Laplace approximations (cf. Rue et al. (2009)) might be used instead. We intend to explore both alternatives in a follow up paper.

Appendix A. Index obtained from Model 3

Table 3. Montly index. Base Jan, 1, 2005 = 100.

Date		Index	Date		Index
Nov	2004	98.83	Mar	2008	115.91
Dec	2004	99.51	Apr	2008	115.02
Jan	2005	100.45	May	2008	114.22
Feb	2005	101.33	Jun	2008	113.59
Mar	2005	102.19	Jul	2008	113.08
Apr	2005	103.09	Aug	2008	112.55
May	2005	104.00	Sep	2008	111.91
Jun	2005	104.99	Oct	2008	111.08
Jul	2005	106.08	Nov	2008	110.13
Aug	2005	107.33	Dec	2008	109.18
Sep	2005	108.67	Jan	2009	108.28
Oct	2005	110.10	Feb	2009	107.51
Nov	2005	111.60	Mar	2009	106.85
Dec	2005	113.17	Apr	2009	106.34
Jan	2006	114.80	May	2009	106.08
Feb	2006	116.34	Jun	2009	106.03
Mar	2006	117.84	Jul	2009	106.22
Apr	2006	119.29	Aug	2009	106.60
May	2006	120.63	Sep	2009	107.04
Jun	2006	121.82	Oct	2009	107.40
Jul	2006	122.86	Nov	2009	107.56
Aug	2006	123.75	Dec	2009	107.54
Sep	2006	124.48	Jan	2010	107.47
Oct	2006	125.07	Feb	2010	107.46
Nov	2006	125.51	Mar	2010	107.50
Dec	2006	125.80	Apr	2010	107.56
Jan	2007	125.93	May	2010	107.58
Feb	2007	125.85	Jun	2010	107.53
Mar	2007	125.57	Jul	2010	107.44
Apr	2007	125.14	Aug	2010	107.30
May	2007	124.64	Sep	2010	107.14
Jun	2007	124.09	Oct	2010	106.98
Jul	2007	123.46	Nov	2010	106.76
Aug	2007	122.67	Dec	2010	106.45
Sep	2007	121.68	Jan	2011	105.97
Oct	2007	120.55	Feb	2011	105.31
Nov	2007	119.46	Mar	2011	104.47
Dec	2007	118.46	Apr	2011	103.45
Jan	2008	117.55	May	2011	102.86
Feb	2008	116.73			

References

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2003), *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC.

Baranzini, A., Ramirez, J., Schaerer, C. and Thalmann, P., eds (2008), *Hedonic Methods in Housing Markets*, Springer Verlag, New York, NY.

Biggeri, L. and Guido Ferrari, eds (2010), *Price Indexes in Time and Space*, Springer Verlag.

Borst, R. (2008), 'Time-Varying Model Parameters: Obtaining Time Trends in a Hedonic Model Without Specifying Their Functional Form', *Journal of Property Tax Assessment & Administration* **6**(4), 29–36.

Bourassa, S., Hoesli, M. and Sun, J. (2006), 'A simple alternative house price index method', *Journal of Housing Economics* **15**(1), 80–97.
URL: *http://linkinghub.elsevier.com/retrieve/pii/S1051137706000064*

Brian J. Smith, Jun Yan, M. K. C. (2008), 'Unified geostatistical modeling for data fusion and spatial heteroskedasticity with R package ramps', *Journal of Statistical Software* **25**(10), 1–21.

Can, A. and Megbolugbe, I. (1997), 'Spatial Dependence and House Price Index Construction', *Journal of Real Estate Finance and Economics* **222**, 203–222.

Clapp, J. M. (2004), 'A Semiparametric Method for Estimating Local House Price Indices', *Real Estate Economics* **32**(1), 127–160.
URL: *http://doi.wiley.com/10.1111/j.1080-8620.2004.00086.x*

Court, A. (1939), *Hedonic Price Indexes with Automotive Examples*, General Motors Corporation, pp. 99—-117.

Cressie, N. and Wikle, C. K. (2011), *Statistics for Spatio-Temporal Data*, Wiley.

Departamento de Vivienda Obras Públicas y Transportes (2011), 'Eskaintza inmobiliarioa. Oferta inmobiliaria'.
URL: *http://www.euskadi.net*

Dubin, R. (1988), 'Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms', *The Review of Economics and Statistics* **70**(3), 466–474.

Fotheringham, S., Charlton, M. and Brunsdon, C. (2002), *Geographically Weighted Regression: The Analysis of Spatially Varying Relationships*, Wiley.

García Montalvo, J. (2007), '¿Cuánto aumenta realmente el precio de la vivienda?', *Indice. Revista de Estadística y Sociedad* **22**, 10–11.

Gelfand, A. E., Diggle, P. J., Fuentes, M. and Guttorp, P., eds (2010), *Handbook of Spatial Statistics*, CRC Press.

Gelfand, A., Ecker, M., Knight, J. and Sirmans, C. (2004), 'The Dynamics of Location in Home Price', *The Journal of Real Estate Finance and Economics* **29**(2), 149–166.

Geniaux, G. and Napoléone, C. (2008), *Semi-Parametric Tools for Spatial Hedonic Models: An Introduction to Mixed Geographically Weighted*

Regression and Geoadditive Models, in Baranzini et al. (2008), pp. 101–127.

Hastie, T. J. and Tibshirani, R. J. (1991), *Generalized Additive Models*, 2nd. edn, Chapman & Hall, London.

Henneberry, J. (1998), 'Transport Investment and House Prices', *Journal of Property, Valuation and Investment* **16**(2), 144–157.

INE (2009), Índice de Precios de Vivienda. Metodología, Technical report, Instituto Nacional de Estadística (INE).
URL: *http://www.ine.es*

Kestens, Y., Thériault, M. and Des Rosiers, F. (2005), 'Heterogeneity in hedonic modelling of house prices: looking at buyers' household profiles', *Journal of Geographical Systems* **8**(1), 61–96.

Kryvobokov, M. and Wilhelmsson, M. (2007), 'Analysing location attributes with a hedonic model for apartment prices in Donetsk, Ukraine', *International Journal of Strategic Property Management* pp. 157–178.

LeSage, J. and Pace, R., eds (2011), *Advances in Econometrics*, Emerald, New York, NY.

Ministerio de la Vivienda (retrieved 10-June-2010), Notas metodológicas. Undated report.
URL: *http://www.mviv.es/es/pdf/otros/NMPV.pdf*

McMillen, D. P. (2011), *Employment subcenters and home price appreciation rates in metropolitan Chicago*, in LeSage and Pace (2011), pp. 237–257.

McMillen, D. and Redfearn, C. (2007), Estimation, Interpretation, and Hypothesis Testing for Nonparametric Hedonic House Price Functions.

Meese, R. A. and Wallace, N. E. (1997), 'The Construction of Residential Housing Price Indices: A Comparison of Repeat-Sales, Hedonic-Regression, and Hybrid Approaches', *Journal of Real Estate Finance and Economics* **73**, 51–73.

Pace, R., Barry, R., Gilley, O. and Sirmans, C. (2000), 'A method for spatial-temporal forecasting with an application to real state prices', *International Journal of Forecasting* **16**, 229–246.

Robinson, P. (1988), 'Root-n-consistent semiparametric regression', *Econometrica* **56**, 931–954.

Rodríguez López, J. (2007), 'Los Índices de precios de la vivienda. Problemática.', *Índice. Revista de Estadística y Sociedad.* **22**, 14–16.

Rosen, S. (1974), 'Hedonic prices and implicit markets: Product differentiation in pure competition', *The Journal of Political Economy* **82**, 34–55.

Rue, H., Martino, S. and Chopin, N. (2009), 'Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2), 319–392.
URL: *http://doi.wiley.com/10.1111/j.1467-9868.2008.00700.x*

Stock, J. (1989), 'Nonparametric policy analysis', *Journal of the American Statistical Association* **84(406)**, 567–575.

Vidaurre, D., Bielza, C. and Larrañaga, P. (2011), 'Lazy lasso for local regression', *Computational Statistics* . To appear.
URL: *http://www.springerlink.com/index/10.1007/s00180-011-0274-0*

Vogt, A. and Barta, J. (1996), *The Making of Tests for Index Numbers*, Physica-Verlag.

Wheeler, D. C. (2006), Diagnostic tools and remedial methods for collinearity in linear regression models with spatially varying coeffcients, PhD thesis, Ohio State University.

Wheeler, D. and Tiefelsdorf, M. (2005), 'Multicollinearity and correlation among local regression coefficients in geographically weighted regression', *Journal of Geographical Systems* **7**(2), 161–187.
URL: *http://www.springerlink.com/index/10.1007/s10109-005-0155-6*

Wood, S. (2001), '`mgcv`: GAMs and generalized ridge regression for R', *R News* **1**(2), 20–25.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Wood, S. (2004), 'Stable and efficient multiple smoothing parameter estimation for generalized additive models', *Journal of the American Statistical Association* **99**, 673–686.

(M.B. Bárcena and F. Tusell) Departamento Estadística y Econometría, Universidad del País Vasco, Bilbao (Spain)

(M.B. Palacios) Departamento Estadística e I.O., Universidad Pública de Navarra, Pamplona (Spain)

(P. Menéndez) University of New South Wales, Sydney (Australia)

E-mail address: `fernando.tusell@ehu.es`