Noname manuscript No.
(will be inserted by the editor)

Housing price indexes for small areas

Smoothing in time and space

M.J. Bárcena · M.C. González-Morgado · P. Menéndez-Galvan, · F. Tusell

Received: date / Accepted: date

Abstract This paper presents yet another alternative to estimate house price indexes and their evolution over time and space. The model presented here builds from Bárcena et al. (2011) Bárcena et al. (2013) which made use of a space-varying quality adjustment fitted by geographically weighted regression (GWR) and a smooth time adjustment common to the whole area analyzed. The model presented here is a natural evolution from the aforementioned work and a refinement in the sense that enables the analyst to estimate different time adjustments for different locations. An application is given showing the different evolution of house prices even in neighbouring districts of the Bilbao urban area.

Keywords Housing prices · price indices · semi-parametric models · GWR

1 Introduction

There have been many modelling proposals for housing prices. In one way or another, they all try to account for quality, location and, where transactions from different periods are analyzed, time.

We thank the Departamento de Medio Ambiente, Planificación Territorial y Vivienda of the Basque Government for permission to use the data. Special thanks are due to Aitor Puerta.

M.J. Bárcena
Facultad de Economía y Empresa, UPV/EHU, Bilbao, Spain
E-mail: mariajesus.barcena@ehu.es

M.C. González-Morgado
Facultad de Economía y Empresa, UPV/EHU, Bilbao, Spain
E-mail: mariacristina.gonzalezm@ehu.es

P. Menéndez
Department of Econometrics and Business Statistics, Monash University, Melbourne, Australia
E-mail: patriciamenendez1@gmail.com

F. Tusell
Facultad de Economía y Empresa, UPV/EHU, Bilbao, Spain
E-mail: fernando.tusell@ehu.es

The usual way to model quality is through a hedonic part which values different attributes of a house: year of construction, type of construction, surface, equipment and facilities, such as swimming pools, lifts, garages, etc.

Location is of paramount importance in the valuation of properties. To some extent it could be accounted for by attributes such as distance to city center, proximity to public transport, recreation facilities, etc. However the use of attributes such as this never exhausts the influence of location, which is usually introduced in the model explicitly. This can be done non-parametrically via smooth functions of geographical coordinates, such as smooth-plate splines, or by letting the coefficients of the hedonic part vary freely over space: geographically weighted regression (GWR) is a popular way of doing so.

The effect of time usually enters models by way of a trend which attempts to account for different prices over time of otherwise identical properties. It can be argued that no property really remains the same over time, if anything else because the environment changes: what used to be an isolated house may in time become the center of a newly developed area, or close to new public transport. Still, the use of trends to account for changes may be a convenient simplification.

The plan of the paper is as follows: Section 2 describes a model which generalizes our previous work localizing estimated trends in prices. Section 4 illustrates its performance on a large data set with over 230,000 observations, extending over the period 2005-2017. Section 5 closes with some comments and conclusions.

2 A semi-parametric model with local trends

Our previous work was concerned with the estimation of price indices for housing. Our raw data consisted of offered prices published by one of the leading property webs in the country.

The specification chosen was,

$$\log(P_{it}) = \sum_{j=1}^{p} \beta_{ij} z_{ij} + s(t) + \varepsilon_{it} \tag{1}$$

In expression (1), P_{it} is the price per square meter at time t of a house located at coordinates[1] (x_i, y_i). The observed value of attribute j for the house at said location is given by z_{ij}; it can be qualitative (e.g. existence of central heating) or numeric (e.g. total surface). The hedonic coefficients β_{ij} give the valuation of attribute j at location (x_i, y_i). Finally, $s(t)$ is a smooth function of time capturing the evolution of prices. Since we targeted small areas (the city of Bilbao, Spain; see for instance Bárcena et al. (2011, 2013)), it made sense to consider a single trend over the whole area.

It is straightforward to implement a back-fitting estimation routine (cf. Hastie and Tibshirani (1991), § 4.4) for the model in (1), once we have routines for the different tasks. These are readily available in, for instance, the R language, R Core Team (2018). A GWR routine (see Harris et al. (2010)) is available in package `spgwr`, Bivand and Yu (2017)). A smoothing spline routine (see for instance Eubank

[1] As the area we will work with can be well approximated by a flat surface, projected UTM coordinates are used.

(1988); Hastie and Tibshirani (1991)) can be obtained from the `mgcv` package, Wood (2017)).

The procedure can be sketched as follows:

Algorithm 1: Global trend with spatially varying attribute effects

Data: For $i \in I$, $t \in T$: price of house i at time t, P_{it};
For $i \in I$, $j \in J$: attribute j for house i, z_{ij};
For $i \in I$: coordinates of house i, (x_i, y_i);
Preset tolerance η.
Result: Estimated values of β_{ij} and $s(t)$.
Set $s^{(0)}(t) = 0$, $k = 1$; 1
while $\max_t |s^{(k)}(t) - s^{(k-1)}(t)| > \eta \cdot \max_t |s^{(k)}(t)|$ **do** 2
Use GWR to fit $[\log(P_{it}) - s^{(k-1)}(t)] = \sum_{j=1}^{p} \beta_{ij}^{(k-1)} z_{ij} + \varepsilon_{it}^{(k)}$; 3
Compute residuals $\hat{\varepsilon}_{it}^{(k)} = \log(P_{it}) - \sum_{j=1}^{p} \hat{\beta}_{ij}^{(k-1)} z_{ij}$; 4
Smooth residuals $\hat{\varepsilon}_{it}^{(k)}$ over time to compute $s^{(k)}(t)$; 5
$k = k + 1$; 6
end 7
return $\hat{\beta}_{ij}^{(k)}$, $s^{(k)}(t)$ as final estimates of β_{ij} and $s(t)$. 8

The back-fitting algorithm iterates lines 2 to 7 in Algorithm 1, estimating the parametric in alternation with the non-parametric part until convergence; the loop is exited when two successive estimates of the non-parametric time trend differ by less than a fraction η of the largest value.

In line 5 of Algorithm 1, a smoothing spline is used; $s^{(k)}(t)$ is estimated as the piecewise cubic polynomial $g(t)$ which minimizes

$$\sum_{i \in I} (\hat{\varepsilon}_{it}^{(k)} - g(t))^2 + \lambda \int (g''(t))^2 dt \tag{2}$$

Alternatives to the algorithm above exist. For instance, Brunauer et al. (2012) propose a Generalized Additive Model (GAM) specified as:

$$\eta_i = f_1(z_{i1}) + f_2(z_{i2}) + \ldots + f_q(z_{iq}) + x_i{}'\gamma \tag{3}$$

where $x_i{}'\gamma$ is the parametric part which is considered global and $f_i(z_{ik})$ are smooth functions of time or spatial indices to accommodate variation in time and/or space. In our model, instead, space variation is accounted for by GWR estimation, and only one smooth function (of time) is used to account for time variation. Many more proposals have been put forward: articles with recent literature reviews include Copiello (2020) and P.Gargallo et al. (2017).

The whole area can of course be divided in smaller regions to apply the previous approach and obtain price trends $s(t)$ for each sub-region. However, a different strategy is followed that allows for local trends while still borrowing strength from neighbouring areas. If we want to estimate a local trend around location (x_ℓ, y_ℓ) we can replace $g(t)$ computed as the minimizer of (2) by the minimizer $g_\ell(t)$ of

$$\sum_{i \in I} w_{i\ell} (\hat{\varepsilon}_{it}^{(k)} - g_\ell(t))^2 + \lambda \int (g_\ell''(t))^2 dt \tag{4}$$

were $w_{i\ell}$ are coefficients which weight more residuals $\hat{\varepsilon}_{it}^{(k)}$ which are closer to the calibration point ℓ; this produces for each point in space a different price trend. A common specification of $w_{i\ell}$ is:

$$w_{i\ell} = \phi(|x_i - x_\ell|^2 + |y_i - y_\ell|^2; b) \tag{5}$$

where ϕ is any kernel function (Gaussian, triangular, bisquare...) and b is a parameter controlling bandwidth. It should be noted that b need not be coincident with the bandwidth used in the GWR.

Using (4) instead of (2) in line 5 of Algorithm 1 produces a derivative algorithm which yields estimates around the spatial point (x_ℓ, y_ℓ). Data points near the calibration point of coordinates (x_i, y_i) are more heavily weighted, so the estimated spline for location reflects primarily the evolution of prices nearby. Although the differences with Algorithm 1 are conceptually small, we give the detailed description in Algorithm 2.

Algorithm 2: Both time trend and attribute effects spatially varying

Data: For $i \in I$, $t \in T$: price of house i at time t, P_{it};
For $i \in I$, $j \in J$: attribute j for house i, z_{ij};
For $i \in I$: coordinates of house i, (x_i, y_i);
For $\ell \in L$: coordinates (x_ℓ, y_ℓ) of locations at which trends are computed;
Preset tolerance η.
Result: Estimates, list of estimated values of β_{ij} and $s(t)$ for each ℓ.
foreach ℓ *in* L **do** 1
  Set $s^{(0)}(t) = 0$, $k = 1$; 2
  For $i \in I$, compute weights $w_{i\ell} = \phi(|x_i - x_\ell|^2 + |y_i - y_\ell|^2)$; 3
  while $\max_t |s^{(k)}(t) - s^{(k-1)}(t)| > \eta \cdot \max_t |s^{(k)}(t)|$ **do** 4
    Use GWR to fit $[\log(P_{it}) - s^{(k-1)}(t)] = \sum_{j=1}^{p} \beta_{ij}^{(k-1)} z_{ij} + \varepsilon_{it}^{(k)}$; 5
    Compute residuals $\hat{\varepsilon}_{it}^{(k)} = \log(P_{it}) - \sum_{j=1}^{p} \hat{\beta}_{ij}^{(k-1)} z_{ij}$; 6
    Smooth weighted residuals $w_{i\ell}\hat{\varepsilon}_{it}^{(k)}$ over time to compute $s^{(k)}(t)$; 7
    $k = k + 1$; 8
  end 9
  Estimates[ℓ] = list($\hat{\beta}_{ij}^{(k)}$, $s^{(k)}(t)$) 10
end 11
return Estimates. 12

3 Related work

There is a vast literature on spatio-temporal models. For a good, book-length introduction, see Cressie and Wikle (2011). LeSage and Pace (2004) contains also a number of contributions. We cannot undertake a comprehensive review and will only comment on work whose motivation or method is closest to ours.

4 An example of use

4.1 Motivation

Algorithm 1 was successfully used to compute a price index for the city of Bilbao (Spain), Bárcena et al. (2011), Bárcena et al. (2013) and has also been used by other authors, e.g. Widłak et al. (2015, 2017).

It was observed in our previous work that, contrary to our assumption that the trend of prices could be assumed unique for all units within a relatively small area (such as the city of Bilbao), this was not so: some areas appear to have weathered the burst of the housing prices bubble much better than others. In fact, areas in which different price trends exist appear to be often the case with housing markets. For a literature review on housing market segmentation see Helbich et al. (2013).

The obvious answer of fitting a model to each area appearing to display a different price evolution is not always feasible, as the areas may be small and the available sample in them insufficient to support the required computation. Algorithm 2 has been conceived instead as a way to fit price trends (and compute indices) for very small areas while still "gathering strength" from information in the vicinity. It extends to the estimation of the time trend the same idea that GWR implements to estimate the attribute effects; spatial weighting, which, unlike in Algorithm 1, is now also applied to the residuals computed in line 6 of Algorithm 2.

4.2 Data

A data set of prices of dwellings on sale has been obtained from one of the leading web portals. These data extends from 2005 to 2017 and covers all the Basque Country, with 237,878 observations after discarding non geocoded or otherwise unusable data. It is thus much more comprehensive (and detailed) than the data set used in our previous work Bárcena et al. (2013, 2014).

However, these observations are very unevenly distributed with the bulk of it in the three largest towns. For the example next only data from 2011 to 2017 has been used and only for the metropolitan area of Bilbao, as only there data density in time and space supports the use of Algorithm 2. For the illustration presented below, the number of effective observations is just over 30,000.

Figure 1 shows a heat map the Greater Bilbao area, a conurbation housing over 1 million people along both banks of the Nervión river. Color coded are smoothed median prices in euros per square meter in 2012 and 2017. It is apparent that the city of Bilbao itself and the city of Getxo (label Algorta, right bank) are the most highly priced areas. It is also apparent that there has been a general median price drop in the five years from 2012 to 2017, visible in the fainter tones of red, which in places turn to blue.

When we look in closer detail to each of the two areas, another feature is apparent: the price drop does not appear to have been uniform all over space. In Figure 2 we see a clear shrink of the reddish portion of central Bilbao (or deeper blue tones in the peripheral districts of the city) in 2017 relative to 2012, reflecting the drop of prices between these two dates. However, prime locations in the financial district and newly developed areas have been least affected

Fig. 1 Metropolitan area of Bilbao: evolution of median prices in € per m^2 between 2012 and 2017

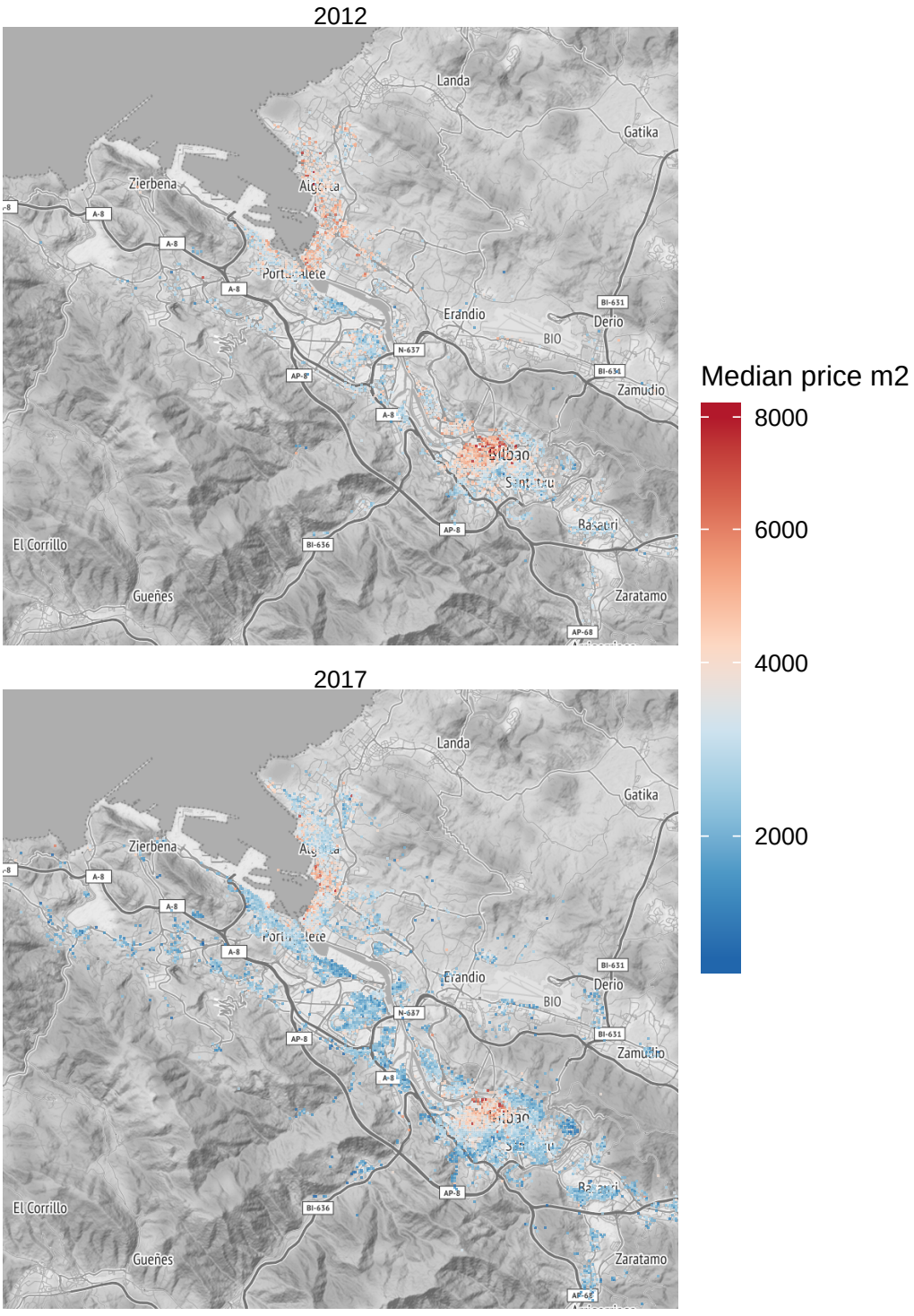


Fig. 2 City of Bilbao: evolution of median prices in € per m^2 between 2012 and 2017

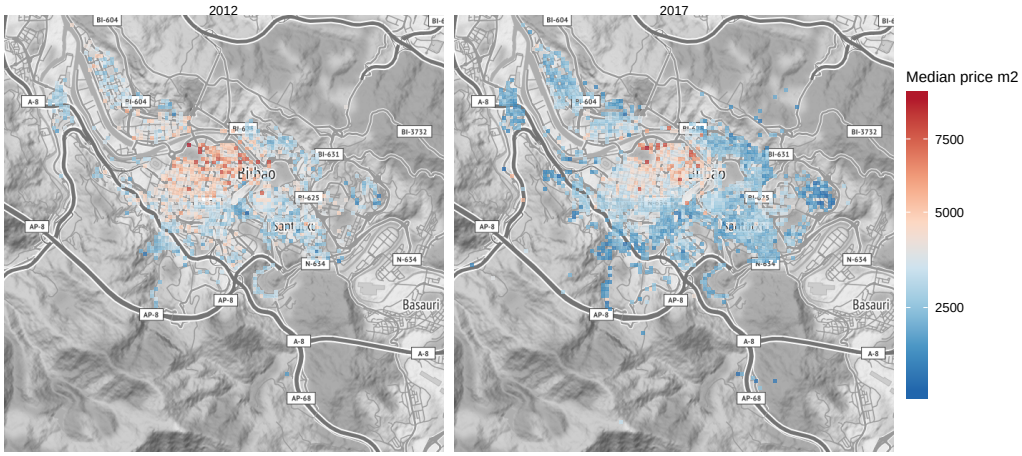
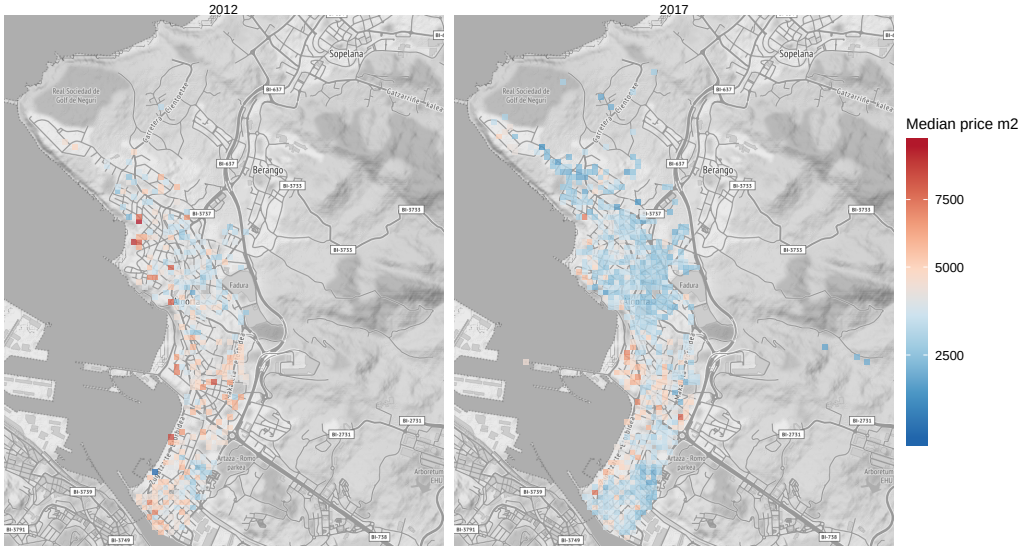


Fig. 3 City of Getxo: evolution of median prices in € per m^2 between 2012 and 2017



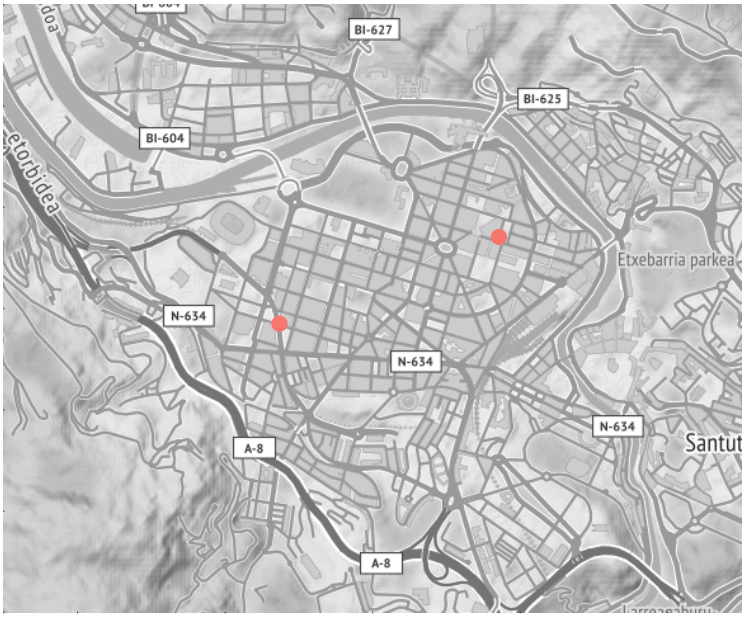
This pattern is even clearer in the case of Getxo, Figure 3. There is a trend towards lower prices per square meter, aparent in fainter red or even a switch from reddish to blue, affecting all but a restricted area at the sea front —a place of very expensive dwellings, seemingly less affected by the price drop. The message both Figures 2 and 3 convey is that price evolution can be quite inhomogeneous even inside a single town; and, in the two cases presented, it seems that prices of the more expensive dwellings are more resilient to a weak market than the rest.

4.3 Computation of local price indices

The computation described in Algorithm 2 has been implemented in an R package[2] of name `ipv`, in function `BackFittingLocal`. All the user has to provide in addition to what is required to fit a single time trend[3] is:

1. A list of locations at which we want the time trend computed. These locations will likely be the centroids of districts of neighbourhoods which we suspect different from the rest, either on *a priori* grounds or after looking at heat maps such as Figures 2 and 3.
2. A bandwidth giving the maximum distance at which residuals computed in line 6 of Algorithm 2 will be given non-zero weight. This bandwidth is unrelated to the one used in the GWR part of the algorithm.

Fig. 4 Central area of Bilbao City. Marked are two locations at which local indices are computed, distant 1297 meters from each other. North East mark is "Plaza Ensanche", central to a high prices area; South West mark is "General Eguía", central to an area more severely hit by price drops.

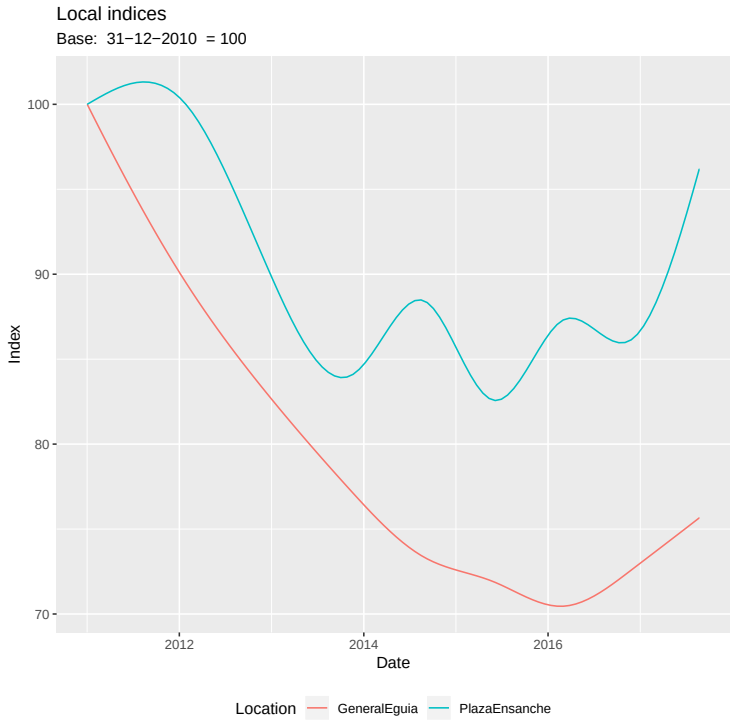


[2] Presently not on CRAN, but available from the authors. Notice, though, that we are prevented from distributing the data, so eventual users will not be able to reproduce the vignette and examples shown. Also, documentation (in Spanish) is not yet translated to English.

[3] Which is the purpose of function `BackFitting` in the same package.

It is important to realize that for each location in 1) a complete back-fitting iteration as described in Algorithm 1 must be run, which is itself a rather heavy computation if the number of observations is large. The computing effort grows linearly with the number of locations selected, and for problems of realistic size such as the example presented will be non trivial. Thankfully, R package `spgwr`, Bivand and Yu (2017), which we use as a building block in Algorithms 1 and 2, provides for parallel computation. For real size problems, a multi-core machine will be almost mandatory.

Fig. 5 Local indices for the two different locations marked in Figure 4. GWR bandwidth is 500 meters, bandwidth for spline weighting is 200 meters, kernel is gaussian.



In the interest of brevity, we will only compute two local indices for two areas of Bilbao, where the density of observations is greater. Figure 4 is a blow-up of the central area of Figure 2, giving a detailed view of downtown Bilbao. The two red dots mark two locations (Plaza Ensanche and General Eguía) around which cluster observations that Figure 2 suggests have experienced a different price evolution. The straight line distance between the two points is 1297 meters.

Algorithm 2 has been used to compute local indices at both locations with results that can be seen in Figure 5. The parametric part estimated by means of GWR (line 5 of Algorithm 2) uses variables such as number of rooms, availability of garden, elevator, terraces or parking space. The bandwidth is set at 500 meters

and a gaussian kernel is used: this implies that if observations right at the red mark are given a weight of 1, those 500 meters away weight 0.6065 and those away 1000 meters weight only 0.1353[4].

On the other hand, the spatial weights for the spline smoother ($w_{i\ell}$ in line 7 of Algorithm 2) are computed using a bandwidth of 200 meters. We emphasize that both bandwidths may be different. The first may be understood as setting the region within which the valuation of attributes in the hedonic part of the model (the parametric model estimated by GWR) is similar; the second, as setting the region whose price evolution in time is similar. Both bandwidths or either one could theoretically be set by using cross-validation, but the computational effort would be very high. In addition, it is usually the case that we are interested in trends for specific areas, which imply the choice of bandwidth. We note however that computationally lighter alternatives have been proposed, Murakami et al. (2019), which make the use of cross-validation feasible.

Regarding the temporal smoothing, we have used 9 equivalent degrees of freedom to compute both local indices. The results can be seen in Figure 5. The base has been set at 31 dec 2010, so the two indices start at the same point. It can be seen that for the first location ("Plaza Ensanche", blue trace) the price drop appears to have ended by late 2013 with some fluctuations afterwards and a clear rebound in 2017. For the second location ("General Eguía", red trace) the price drop continued till early 2016 and the recovery is less marked and leaves the index in mid 2017 clearly below its counterpart for "Plaza Ensanche". Not only this confirms what we could grasp from the heat maps in Figure 2, but also gives additional information on the dynamic of prices.

5 Discussion

An algorithm for the computation of local house price indices has been introduced, a natural evolution of Algorithm 1, and its performance demonstrated.

Among the strengths, it is relatively simple to implement, produces very detailed information and is easy to understand and interpret.

Among the weaknesses, as we see them, it requires very heavy computation. A multi-core machine can be used, but still the computational burden is important for realistic problems, all the more so if one attempts to set bandwidths by cross-validation: there are two spatial bandwidths rather than one —for the GWR part and for the spline— plus the analyst has to decide on the temporal bandwidth, which is governed by the choice of degrees of freedom in the spline.

When indices for only a restricted area are sought, such as in the example presented, one trick that speeds computations tremendously is to discard all observations which are sufficiently away from the locations of interest: "sufficiently away" can be something like four times the largest bandwidth. An observation 4 bandwidths away receives already a weight of only 0.000335 with the gaussian kernel that we used, so all observations beyond can be safely neglected in the local computations.

Aspects that need elaboration, but are not specific to the algorithm presented, concern the anisotropy of the space. This is particularly relevant when dealing

[4] The weight function `gwr.Gauss` in the R package `spgwr`, Bivand and Yu (2017), is used.

with small areas rather than averaging over wide regions. It is often the case that areas with different behaviour are limited by administrative boundaries or geographical features, such as rivers: one might want to compute local indices taking into account these facts, which an isotropic kernel such as we have used neglects. One can consider complementing the kernel with qualitative variables (such as "left bank", "right bank", "district x") to further restrict the scope of observations which are used in the computation of local indices, but this is of necessity an ad-hoc solution which needs re-implementing for each particular case.

References

Bárcena M, Menéndez P, Palacios M, Tusell F (2011) Measuring the Effect of the Real Estate Bubble: a House Price Index for Bilbao. BILTOKI 2011.07, University of the Basque Country (UPV/EHU), URL http://hdl.handle.net/10810/5463

Bárcena M, Menéndez P, Palacios M, Tusell F (2013) Data Mining Applications in R, Academic Press, chap A Real-Time Property Value Index based on Web Data

Bárcena M, Menéndez P, Palacios M, Tusell F (2014) Alleviating the effect of collinearity in geographically weighted regression (GWR). Journal of Geographical Systems 16:441–466

Bivand R, Yu D (2017) spgwr: Geographically Weighted Regression. URL https://CRAN.R-project.org/package=spgwr, r package version 0.6-32

Brunauer W, Feilmayr W, Wagner K (2012) A new residential property price index for Austria. Statistiken (Q3):90–102

Copiello S (2020) Spatial dependence of housing values in northeastern italy. Cities

Cressie N, Wikle CK (2011) Statistics for Spatio-Temporal Data. Wiley

Eubank RL (1988) Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York

Harris P, Fotheringham A, Crespo R, Charlton M (2010) The Use of Geographically Weighted Regression for Spatial Prediction: An Evaluation of Models Using Simulated Data Sets. Mathematical Geosciences 42(6):657–680, DOI 10.1007/s11004-010-9284-7

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

Helbich M, Brunauer W, Hagenauer J, Leitner M (2013) Data-driven regionalization of housing markets. Annals of the Association of American Geographers 103(4):871–889, DOI 10.1080/00045608.2012.707587

LeSage J, Pace RK (eds) (2004) Spatial and Spatiotemporal Econometrics, vol 18. Emerald

Murakami D, Tsutsumida N, Yoshida T, Nakaya T, Lu B (2019) Scalable GWR: A linear-time algorithm for large-scale geographically weighted regression with polynomial kernels. arXiv (1905.00266), URL http://arxiv.org/abs/1905.00266, 1905.00266

PGargallo, JAMiguel, MJSalvador (2017) MCMC Bayesian spatial filtering for hedonic models in real estate markets. Spatial Statistics pp 47–67

R Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL `https://www.R-project.org/`

Widłak M, Waszczuk J, Olszewski K (2015) Spatial and hedonic analysis of house price dynamics in Warsaw. Tech. Rep. NBP Working Papers 197, Narodowy Bank Polski. Economic Research Department

Widłak M, Waszczuk J, Olszewski K (2017) Spatial and hedonic analysis of house price dynamics in Warsaw, Poland. Journal of Urban Planning and Development 143

Wood S (2017) Generalized Additive Models: An Introduction with R, 2nd edn. Chapman and Hall/CRC