Computers, Environment and Urban Systems 37 (2013) 59–69

Contents lists available at SciVerse ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/compenvurbsys

HouSI: Heuristic for delimitation of housing submarkets and price homogeneous areas

V. Royuela [a,*], Juan C. Duque [b]

[a] Universitat de Barcelona, Grup d'Analisi Quantitativa Regional AQR-IREA, Barcelona, Spain
[b] Research in Spatial Economics (RiSE group), Department of Economics, EAFIT University, Carrera 49 7 Sur – 50. Medellin, Colombia

ARTICLE INFO

Article history:
Received 10 June 2011
Received in revised form 27 March 2012
Accepted 7 April 2012
Available online 15 August 2012

Keywords:
Housing submarket segmentation
Hedonic prices
Spatial clustering

ABSTRACT

This paper seeks to address the problem of the empirical identification of housing market segmentation, once we assume that submarkets exist. The typical difficulty in identifying housing submarkets when dealing with many locations is the vast number of potential solutions and, in such cases, the use of the Chow test for hedonic functions is not a practical solution. Here, we solve this problem by undertaking an identification process with a heuristic for spatially constrained clustering, the "Housing Submarket Identifier" (HouSI). The solution is applied to the housing market in the city of Barcelona (Spain), where we estimate a hedonic model for fifty thousand dwellings aggregated into ten groups. In order to determine the utility of the procedure we seek to verify whether the final solution provided by the heuristic is comparable with the division of the city into ten administrative districts.

1. Introduction

A housing market is the collection of alternative locations considered by households as location substitutes, with houses in different submarkets being imperfect substitutes within the same broader market.[1] Housing market segmentation is widely recognised in the literature, and several studies have stressed its importance. Goodman and Thibodeau (2007) propose a number of reasons as to why an understanding of how metropolitan areas are partitioned into housing submarkets is important: (a) it increases the accuracy of statistical models when estimating house prices (Goodman & Thibodeau, 2003); (b) it improves the modelling of spatial and temporal variations in house prices; (c) it improves the ability to price the risk associated with financing homeownership for lenders and investors; (d) it can reduce search costs for those demanding housing; and (e) it avoids inducing spatially correlated errors that bias the coefficients on variables correlated with the errors. Bourassa, Cantoni, and Hoesli (2007) further emphasise the importance of submarket differentiation in their study of house price prediction in a mass appraisal context. They report equally, or even more, accurate predictions when using a traditional hedonic model incorporating a series of dummy variables to define submarkets as when using more complicated models (such as lattice SAR and CAR or geostatistical models).

Many techniques have been adopted in the detection of submarkets and housing segmentation: cluster analysis (Bourassa, Hamelink, Hoesli, & MacGregor, 1999), GIS and ESDA analysis (Tu, Sun, & Yu, 2007), cointegration analysis (Jones et al., 2003), fuzzy clustering (Hwang & Thill, 2009), non-parametric smoothing and spline functions (Pavlov, 2000), neural networks (Kauko, 2004), classification regression trees (Fan, Ong, & Koh, 2006), and household mobility patterns (Jones, Leishman, & Watkins, 2004; Royuela & Vargas, 2009). The most frequently applied technique by far has been the use of hedonic models for house prices. The basic assumption is that hedonic coefficients for housing characteristics, such as living space, capitalise neighbourhood amenities, such as public education. Separate models for global housing markets and for housing submarkets are computed and, then, *F*-tests for nested models are used. These tests tell us whether or not there is a significant reduction in the sum of squared residuals by splitting the subsample into submarkets. The usual result is that housing submarkets matter and this is what is to be expected, especially when using large datasets. In this study, we adopt this approach, considering as our crucial criterion the proximity between the vectors of hedonic price characteristics of each submarket.

The main problem faced by researchers is how best to combine a large number of spatial units into a smaller number of housing submarkets, considering that the number of different ways that

* Corresponding author.
E-mail address: vroyuela@ub.edu (V. Royuela).

[1] Housing submarkets can be the result of several constraints in the spatial arbitrage process within a Housing Market Area: transaction costs, search costs, imperfect information and inelastic supply. In any case, houses in different submarkets are still substitutes, although imperfect. Jones, Leishman, and Watkins (2003) have demonstrated that these differences can be stable over time.

0198-9715/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compenvurbsys.2012.04.005

N dwellings or areas can be grouped into *M* submarkets is particularly large (Cliff & Hagget, 1970; Cliff, Haggett, Ord, Bassett, & Davies, 1975; Keane, 1975).[2] The literature has tackled this problem in two ways. On the one hand, different methods have been used to delineate submarkets (see above). On the other hand, the literature employs a small number of previously defined (in administrative terms, for instance) submarket regions, which might include districts or census tracks (as in Watkins, 2001,[3] Adair, Berry, & McGreal, 1996; Goodman & Kawai, 1982; Schnare & Struyk, 1976). Both approaches provide *discrete* submarkets, which has been criticised on the grounds that neighbourhoods (housing submarkets) are not easily defined *a priori* and, consequently, continuous solutions are deemed preferable (Redfearn, 2009; Sunding & Swoboda, 2010).

Our paper seeks to address the problem of the empirical identification of housing market segmentation, once we assume that submarkets exist. We undertake this identification process by implementing a heuristic for spatially constrained clusters, the "Housing Submarket Identifier" (HouSI). This heuristic combines two strategies: (a) the construction of initial feasible solutions based on the region-growing strategy (Taylor, 1973 and Openshaw, 1977), and (b) the local improvement of a feasible solution based on the Tabu Search algorithm (Openshaw & Rao, 1995 and Duque, Anselin, & Rey, 2011). A more formal definition of the problem of aggregating *N* areas into *M* regions, while optimising a predefined aggregation criterion, is available in Duque, Ramos, and Surinach (2007) and Duque, Church, and Middleton (2011).

HouSI builds on the existing literature on modelling housing submarkets. It achieves this by using estimates of hedonic parameters and their standard errors, which are then incorporated into the objective function to assess the solution quality. Moreover, HouSI can improve the use of hierarchical linear modelling (Goodman & Thibodeau, 1998) as its results depend neither on the starting point nor on the order of adjacency.

It is our belief that the use of discrete housing submarkets rather than continuous delineations of housing submarket differences (built for instance with local weighted regressions)[4] affords the following advantages: (a) it helps policy makers to better define place-based policies; (b) it improves political decisions by targeting the right space to be controlled (Jenkins, 1978) and by ensuring that the specific public policy actions implemented in a region have a homogeneous impact throughout that region (Fischer, 1980); (c) it is easy to implement and understand; and (d) in mass appraisal contexts, the consumption of degrees of freedom is not a great problem.

We examine housing market segmentation by conducting a case study in the city of Barcelona (Spain), where we build a small number of submarkets starting from 40 postal districts. As Hwang and Thill (2009) argue, the body of literature has leaned toward testing the distinctiveness of housing submarkets given *a priori* housing submarkets. Consequently, in order to determine the utility of the procedure we seek to verify whether the final solution provided by the heuristic is comparable with the division of the city into ten administrative districts (as in Adair et al., 1996; Goodman & Kawai, 1982; Schnare & Struyk, 1976). Our results show that the final division of the city undertaken with HouSI is superior to that of the administrative districts identified in terms of housing prices by the Chow test, and that it is statistically significantly better than any random aggregation of housing submarkets.

For simplicity's sake, we do not undertake a review of the literature (for a survey of housing submarkets and related issues, see Watkins, 2001; Kauko, 2004; Goodman and Thibodeau, 2007; Páez, 2009; and Islam & Asami, 2009), but focus rather on presenting the heuristic (Section 2), the housing price hedonic models and case study (Section 3), the results (Section 4), and the main conclusions of our work (Section 5).

2. HouSI: Housing Submarket Identifier

We consider housing submarkets to be partitions of an entire housing market, namely, a city. Dwellings in different submarkets are poor substitutes. We proxy these submarket partitions by defining regions composed of spatially contiguous areas in which the houses are similar in terms of a given set of properties. This assumption of spatial contiguity in the defining of the regions is not arbitrary:

- Government policies and private sector marketing strategies are usually geographically targeted. The use of homogeneous geographic regions to define the applicability and scope of a policy or marketing strategy will increase the probability of achieving the intended effects and of better predicting the unintended effects (Jenkins, 1978).
- Tobler's first law of geography[5] (Tobler, 1970) suggests that unobserved urban structures in the data, as well as unobservable human associations (ethnic, family ties, neighbourhood interactions, etc.), are likely to show geographic patterns that can be bounded by spatially contiguous regions.

The aggregation of a set of geographic areas into spatially contiguous regions while optimising an aggregation criterion has been referred to in the literature by various names: region-building (Byfuglien & Nordgard, 1973), conditional clustering (Lefkovitch, 1980), clustering with relational constraints (Ferligoj & Batagelj, 1982), constrained clustering (Legendre, 1987), contiguity constrained clustering (Murtagh, 1992), regional clustering (Maravalle & Simeone, 1995), contiguity constrained classification (Gordon, 1996, 1999), regionalization (Wise, Haining, & Ma, 1997), or clustering under connectivity constraints (Hansen, Jaumard, Meyer, Simeone, & Doring, 2003).[6] These contributions focus primarily on identifying efficient ways to control for spatial contiguity, formulating different aggregation criteria, and designing strategies for exploring the solution space in search of a near optimal solution. Spatially constrained clustering has been applied to a wide range of empirical problems including: electoral districting (Williams, 1995; Yamada, 2009), school districting (Caro, Shirabe, Guignard, & Weintraub, 2004), sales districting (Ríos-Mercado & Fernández, 2009; Zoltners & Sinha, 1983), health care districting (Pezzella, Bonanno, & Nicoletti 1981), electrical power districting (Bergey, Ragsdale, & Hoskote, 2003), neighbourhood structure definition (Weeks, Hill, Stow, Getis, & Fugate, 2007), Bayesian smoothing techniques (Li, 2007), intra-urban inequality assessment (Weeks, Hill, Getis, & Stow, 2006), unemployment space–time changes (Duque, Artis, & Ramos, 2006), wildness (Comber et al., 2010), among others.

In this paper, we propose a Housing Submarket Identifier (HouSI) heuristic. To the best of our knowledge, this is the first time a heuristic of this type has been applied to the delimitation of hous-

[2] The number of feasible solutions when aggregating *N* areas into *M* spatially contiguous regions is a large number that depends not only on parameters *N* and *M*, but also on the spatial distribution of the areas to be aggregated. Keane (1975) estimated that the number of solutions of aggregating $N = 10$ areas into $M = 5$ regions is between 126 and 42,525.

[3] Watkins proposes a list of complementary alternatives for building housing submarkets: spatial, structural, demander based submarkets. The criticisms that follow would obviously apply to the spatial dimension.

[4] In local weighted regressions model parameters vary in space so as to reflect spatial heterogeneity and, consequently, parameter estimates are a function of the "local" data.

[5] "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236)

[6] See Murtagh (1985), Gordon (1996) and Duque et al. (2007) for a literature review of these methods.

ing submarkets. Below, we describe the overall strategy and the three main components of HouSI.

2.1. Overall strategy

Our starting point is a large set of dwellings (L), which are distributed over a set of areas ($n = 1,\ldots,N$). We need a large number of dwellings per area ($L_1,\ldots,L_N$) in order to estimate a hedonic model of housing prices for every area n, which generates a vector of parameters B_n, and its respective variance–covariance matrix, $V(B_n)$. Our aim is to merge the N areas into M ($m = 1,\ldots,M$) analytical regions, so that each region m contains areas with similar characteristics in terms of hedonic housing prices (B_m) and variance–covariance matrices, $V(B_m)$.

The construction of analytical regions requires the definition of an aggregation criterion that evaluates each feasible solution, which can be used to choose the best among all visited solutions. The solution space is explored in two phases: (1) by constructing a set of feasible solutions, where each solution is generated by growing regions from an initial set of M areas selected at random; (2) by applying a local search process that seeks to improve a given feasible solution by exploring its neighbouring feasible solutions while avoiding entrapment in a local optima. By adopting this approach we consider two available search operations to create new solutions in heuristics: (i) exploration, or diversification, in the construction phase, which enables the heuristic to find new zones in the solution space that may contain potentially better solutions; and (ii) exploitation, or intensification, in the local search phase, which generates new solutions by performing small changes in existing solutions.[7]

2.2. Objective function or aggregation criterion

The aggregation criterion is a critical component in each clustering process, varying from application to application, where its purpose is to evaluate the quality of a given candidate solution. Examples of aggregation criteria in the literature include: maximising the level of intraregional homogeneity in terms of a set of socioeconomic variables; minimising the difference in regional population levels between the regions; maximising the level of spatial compactness of the designed regions; maximising a measure of performance of an econometric model, among many others. In addition, various attempts have been made at quantifying concepts such as homogeneity, equality and compactness. Some authors have also proposed aggregation criteria that result from weighted combinations of two or more criteria.[8]

Based on the premise that breaking a housing market into submarkets makes sense as long as the resulting vectors of hedonic parameters are sufficiently different for each submarket (since this indicates that the hedonic characteristics differ depending on the submarket under evaluation), our aggregation criterion seeks to maximise the discrepancies between these submarket vectors of hedonic parameters. In order to find a proper metric to summarise these discrepancies we build a statistic which follows the form of a generalised Wald test (Satorra & Neudecker, 1997):

$$T = NB'_M(H \otimes \overline{Y}^{-})B_M \tag{1}$$

where B_M is the vector including the parameters of all final submarkets (with dimension $M \cdot K \times 1$, M being the number of submarkets and K the number of structural characteristics of the dwellings), $\overline{Y}^{-}$ is the generalised inverse of the average variance–covariance matrix of all subsamples (with dimension $K \times K$), and H is a composite matrix (with dimension $M \times M$) which allows the T statistic to summarise the discrepancy between submarket parameters and the global average, taking into consideration the average variance of the estimates (included in $\overline{Y}^{-}$). Thus, overall, the larger the T statistic, the higher the discrepancy between housing submarkets, and consequently, the solution with the highest T will be preferred.[9] This procedure is theoretically consistent with the definition of housing submarkets whereby dwellings in different submarkets are poor substitutes, and consequently higher discrepancies between submarkets are preferred.[10]

2.3. Construction phase

This component of the heuristic seeks to generate a feasible solution; i.e., to aggregate N small areas into M spatially contiguous analytical regions, or housing submarkets, so that each area is assigned to one and only one submarket, and each submarket comprises at least one area. These solutions are then evaluated with the aggregation criterion to determine the quality of the solution.

Many options are available in the literature for constructing an initial feasible solution, where the choice should take into consideration the following aspects:

- *Shape of the regions*: the shape of the regions depends on the context of application. For example, some solutions require regional compactness for minimising travel distance (in the case of school districting), or minimising the risk of gerrymandering[11] (in the case of electoral districting). Other solutions prefer to allow for irregularly shaped regions so that the resulting regions can capture a wide variety of spatial patterns of socioeconomic variables. As we are unable to make any assumptions regarding the shape of housing submarkets, the possibility of allowing the spatial pattern to dictate the shape of the regions is a key characteristic when deciding on the type of construction method.
- *Capacity of generating a wide range of feasible solutions*: The problem of aggregating N areas into M spatially contiguous regions is classified as being non-deterministic polynomial-time hard (or NP-hard) (Altman, 1997). Here, it is essential that the heuristic is capable of undertaking a good exploration of the solution space so as to increase the possibility of finding a good initial feasible solution and of reducing the chance of premature convergence (Weise, 2009).
- *Speed*: Having a fast algorithm for constructing an initial feasible solution allows us to generate, in a decent amount of time, a large number of feasible solutions from which the best solution can be retained for further improvement.

In keeping with these requirements, we chose to implement the construction phase using the "seeded regions strategy", in which each region starts its growth from an initial area (seed). Subsequently, neighbouring areas are attached to this seed area until all areas are assigned. Selecting the initial set of M seeds at random ensures that each time the construction phase is run, a different feasible solution is provided, which guarantees a good exploration of the solution space. This strategy is also computationally efficient and, unlike other strategies for constructing initial feasible solutions (for example, methods based on location-allocation models),

[7] See Lin and Gen (2009) and Mashinchi, Orgun, and Pedrycz (2011) for more information on the trade-off between these two strategies.

[8] See Johnston (1968), Lankford (1969), and Fischer (1980) for more discussion on the relevance of the aggregation criterion when identifying certain spatial patterns.

[9] Appendix A shows various details of the generalised Wald test.

[10] Royuela and Vargas (2009) apply a similar criterion for finding housing submarkets within a region.

[11] Term used to describe the manipulation of the geographic boundaries of electoral districts in order to benefit a particular party.

it is able to design regions of any shape (compact, elongated, concentric or irregular regions).

During the construction phase, a criterion needs to be defined for selecting the next candidate area to be added to a growing region. A candidate area is any unassigned area that shares a border with at least one growing region. In this case we apply an attribute-based dissimilarity function from each candidate area to the centroid of each growing region where the area can be assigned. The centroid of each growing region (C_g) is calculated as:

$$C_g = \sum_n \sum_i (A_{in} * s_n / S_g) \tag{2}$$

where A_{in} is the value of attribute i (where $i = 1,\ldots,I$, being I the total number of attributes) in area n, s_n is the size (in terms of the number of dwellings) of area n, and S_g is the size of the growing region g, measured as the sum of s_n belonging to the growing region. Thus, the centroid of a given growing region is the size-weighted average of its area attributes.[12] Given the potentially significant size differences between areas (e.g. in terms of population), this weighted-centroid performs better than the simple average.

With the regional centroids, the dissimilarity between a candidate area, i, and each of its candidate adjacent growing regions, j, is calculated with a statistic which follows the form of a Wald test of structural change in the vector of the hedonic price estimates for two subsamples, assuming the existence of different variances in every estimation i and j (Ohtani & Toyoda, 1985; Toyoda & Ohtani, 1986), where B_i and B_j are the vector of parameters of the hedonic regression of region i and j respectively, and $VAR(B_i)$ and $VAR(B_j)$ are the variances of these estimates[13]:

$$W_{ij} = (B_i - B_j)'(VAR(B_i) + VAR(B_j))^{-1}(B_i - B_j) \tag{3}$$

This expression is of particular interest, as the statistic W_{ij} is equal to W_{ji}. This metric helps us to determine the best option for merging an area i with a growing housing submarket (region). The lower the statistic, the lower the discrepancy is between the vector estimates of the hedonic model.[14] This strategy should leave large areas isolated, as the greater the sample size, the smaller the estimation variance will be, and consequently, the W_{ij} statistic can be expected to be higher.

2.4. Local search phase

Once the best initial feasible solution has been selected, the final step in the delineation of housing submarkets involves attempting to improve the initial feasible solution by moving areas between neighbouring regions while seeking to improve the aggregation criterion. This procedure, known as "local search", has been widely applied in spatial aggregation strategies and there exist many heuristics for undertaking it. Some of these are fast, though they can easily become trapped in a local optimal solution (e.g., greedy heuristic); others are computationally expensive, but incorporate strategies that allow them to escape from the local optimal solution. The two most widely recognised heuristics of this kind, within the context of spatial aggregation, are simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) and tabu search (Glover, 1977, 1989, 1990). Recent computational experiments involving spatial aggregation models show that tabu search performs better than simulated annealing in over 95% of cases (Duque, Anselin, et al., 2011).

```
Pseudocode 1: HouSI
M, maxitr, l

1:  $\Lambda$ = set of areas
2:  $\psi = \varnothing$, best initial feasible solution
3:  for $I = 1,2,\ldots,$ matrix do
      CONSTRUCTION PHASE
4:    $\Lambda^u = \Lambda$, set of unassigned areas
5:    $G$ = set of growing regions. It is initiated with $M$ seeds
      selected at random
6:    $\Lambda^u = \Lambda^u - G$
7:    while $\Lambda^u = \varnothing$
8:      $N$ = set of areas that share a border with one or more
        areas in $G$, and $N \subseteq \Lambda^u$
9:      $B$ = area in $N$ that minimises the function $W_{ij}$ (i.e., the
        distance between a given area $i$ and the centroid of the
        growing region $j$, $C_g$), and where area $i$ shares a border
        with growing region $j$
10:     $\Lambda^u = \Lambda^u - \{B\}$, area $B$ is assigned to a neighbouring
        growing region
11:     $G = G \cup \{B\}$
12:   if $T(G) \leqslant T(\psi)$, where $T$ is the generalised Wald test
      (see Eq. (1))
13:     $\psi = G$

LOCAL SEARCH
14: $A = \psi$, aspirational solution
15: $\Phi = \psi$, current solution
16: $c = 1$
17: while $c \leqslant 230\sqrt{M}$
18:   $\eta$ = set with elements $(i,k)$ containing neighbouring
      moves. Thus, for a given feasible solution, a
      neighbouring move is any move of one area, $i$, from its
      current region (the donor region) to another region, $k$,
      (the recipient region), such that this move leads to
      another feasible solution (i.e., it does not break the
      spatial contiguity of the donor region, and the donor
      region contains at least one area after removing area $i$
19:   if $\eta \neq \varnothing$ and $\xi = 0$
20:     $\Omega$ = move in $\eta$ that leads to the lowest value of $T$
21:     $\eta = \eta - \{\Omega\}$
22:   if $\Omega$ is a tabu move
23:     if $T(\Omega) \leqslant T(A)$
24:           $A = T(\Omega)$
25:           $\Phi = T(\Omega)$
26:           $c = 1$
27:           Make the reverse move (i.e. return area $i$ to its
              donor region) tabu, or forbidden, during the
              next $l$ iterations (or moves)
28:     elseif
29:           go to line 18
30:   elseif
31:     if $T(\Omega) > T(A)$
32:           $A = T(\Omega)$
33:           $\Phi = T(\Omega)$
34:           $c = 1$
35:           Make the reverse move (i.e. return area $i$ to its
              donor region) tabu, or forbidden, during the
              next $l$ iterations (or moves)
```

[12] In our case it is applied to both the vector of parameters and the vector of the variance-covariance matrix.

[13] Usually neighboring areas are merged if they pass the classical Chow F-test for nested models. In the housing submarkets literature the Chow test is used to detect whether two submarkets are differentiated, and when the number of observations for every submarket is high, the usual result confirms heterogeneity. In our case, we use the test to merge rather than to separate regions, regardless of the number of observations.

[14] We can merge area 1 with two alternative areas, 2 and 3, and for that purpose we compute the Wald statistic W_{12} and W_{13}. If $W_{12} < W_{13}$ we will merge area 1 with area 2 instead of with area 3.

```
    elseif
        $\Phi = T(\Omega)$
        $c = c + 1$
        Make the reverse move (i.e. return area $i$ to its donor region) tabu, or forbidden, during the next $l$iterations (or moves)
return A
```

See Nagel (1965), Sammons (1978) and Horn (1995) for a review of the different means of generating neighbouring solutions within the context of spatial clustering.

The tabu search heuristic allows for a temporal worsening of the evaluation criterion in the hope of discovering a solution that is better than the "best" solution obtained so far (the *aspirational criterion*). The procedure begins with an initial feasible solution and then moves to the best neighbouring solution, even if this leads to a deterioration in the current aggregation criterion (the *current solution*).[15] To prevent cycles, the reverse move is forbidden (or is tabu) for a predefined number of iterations (*lengthTabu*). A tabu move is allowed only if the move yields a better solution than that provided by the *aspirational criterion*. The heuristic stops when a total of $230\sqrt{M}$ iterations have been performed without any improvement in the aspirational criterion.[16] According to the literature, the most critical parameter in this heuristic is the length of the tabu list, *lengthTabu*. Pseudocode 1 contains a more formal description of the HouSI heuristic.

3. The hedonic model

Wilkinson (1973) constructs a classification of housing characteristics that incorporates structural (house specific) and locational characteristics (neighbourhood specific). Cheshire and Sheppard (1995) use a hedonic model in which both types of variable are used.[17] In most studies it is assumed that housing prices capitalise location characteristics and, consequently, the use of house specific hedonic prices is sufficient to determine whether there are any housing submarkets, whatever the reason might be for their existence (structural, locational, and even demand characteristics, including income and race). When considering housing market segmentation based on pre-existing geographical units, not all studies include locational variables (e.g. Watkins, 2001, merely includes the "crow-fly" distance to the city centre), as the use of dummies for every region captures the spatial specificities.[18]

The empirical model that we estimate here is a logarithmic function, where the log of the housing price depends on the log of the I structural characteristics of the dwellings:

$$\ln(p) = \sum_{i}^{I}\beta_i(\ln(X_i) + \varepsilon \tag{4}$$

Our study is undertaken in the second largest city in Spain, Barcelona, in the north-east of the country. In 2001, the base year for our study, it had a population of 1.5 million inhabitants. The city is divided into 40 postal districts, which are shown in Fig. 1.

Municipal housing price data are drawn from the Spanish Ministry for Housing and refer to the period 2000–2004. The database contains 99,182 dwellings and takes into account the postal district in which the dwelling is located, along with a small number of structural characteristics, including age and size.[19] We do not have, however, the detailed location of every dwelling within every postal district. Had this been available, alternative procedures could have been adopted for building housing submarkets.

Table A1 (Appendix B) shows the basic statistics for all the variables, while Figs. A2.1–A2.3 show the spatial distribution of each variable by postal district. As we seek to identify housing submarkets, we need to use comparable dwellings within the same submarket. Thus, we focus on dwellings that do not present extreme values in terms of price (between 72,000 € and 680,000 €), size (between 40 and 170 m^2), and age (below 35 years old). Consequently, the database is restricted to 50,980 dwellings, i.e. more than 50% of the initial database.[20] On average we have over a thousand dwellings per postal district, with a minimum sample size of over a hundred dwellings per postal district (Postal District – 8007).

4. Results

We estimated a log–log function for our hedonic model in which the structural characteristics (*age* and *size*) were combined with a list of time dummies. Subsequently, we incorporated dummies for the city's postal districts and the interactions with the structural variables of age and size. In order to identify the model's sources of explanatory power, we considered starting with a simple model of structural characteristics (Model 1), subsequently expanded to include time dummies (T_t, in Model 2), local dummies (D_i, in Model 3), and the interactions of these dummies with the structural characteristics (Model 4).

$$\text{Model 1}: \quad \ln(p) = \beta_0 + \beta_1\ln(size) + \beta_2\ln(age) + \varepsilon \tag{5}$$

$$\text{Model 2}: \quad \ln(p) = \beta_0 + \beta_1\ln(size) + \beta_2\ln(age) + \sum_{t=2001}^{2004}\gamma_t T_t + \varepsilon \tag{6}$$

[15] Note that at each iteration, the best neighboring solution does not necessarily lead to an improvement of the aggregation criterion (also known as the *current solution*); however, non-improving moves of this type are allowed for a limited number of consecutive iterations to give the algorithm the capacity to escape from local optimal solutions.

[16] Although the Tabu Search algorithm has been widely applied within the context of spatially constrained clustering, there are only a few studies that evaluate the capacity of the algorithm to navigate the solution space for different parameter values (Blais, Lapierre, & Laporte, 2003; Bozkaya, Erkut, & Laporte, 2003; Ricca & Simeone, 2008). Many papers, including those by Openshaw, do not even mention the number of iterations defined as stopping rule. In our case we decide to use the highest value for the stopping rule reported in the peer reviewed papers that apply this algorithm for spatial clustering. Bozkaya et al. (2003) use 115sqrt(m) and 230sqrt(m); Ricca and Simeone (2008) use 280; and Blais et al. (2003) use 100sqrt(m); where m is the number of regions. As pointed out by Bozkaya et al. (2003), using sqrt(m) instead of m is a common practice that allows us to take into account the problem size in the parameter settings. Since the higher the parameter value for the stopping rule, the higher the possibility of getting a better solution, we chose to use the highest reported value for this parameter: 230sqrt(m).

[17] This approach has been widely used to estimate the value of green areas (Gunn, 2007), forests (Hand, Thacher, & McCollum, 2008), improvements in transportation systems (Yiu & Wong, 2005), public goods (Gravel, Michelangeli, & Trannoy, 2006), among others.

[18] When researchers have spatially disaggregated data, the use of locational variables for every dwelling can be a much more appropriate procedure.

[19] The database employs statistics provided by appraisal firms, as the *real* prices of housing transactions are not published in Spain. We are aware that our procedure presents certain weaknesses, including problems of bias (Dietrich, Harris, & MullerIII, 2000); dispersion (Hansz and Diaz-III, 2003); and econometrics (see Bond & Hwang, 2007, for a list of such problems associated with appraisals). However, our main concern is whether the appraisal firms use a territorially-biased method or procedure in their computations. Fortunately, we understand that this is not the case in Barcelona, where appraisal firms are large and officially accredited. Thus, we assume that price adjustments are not persistently inconsistent in space. This database was previously used in Royuela and Vargas (2009).

[20] We would have worked with a broader final database if we had had access to more structural characteristics. Regrettably, this was not the case.

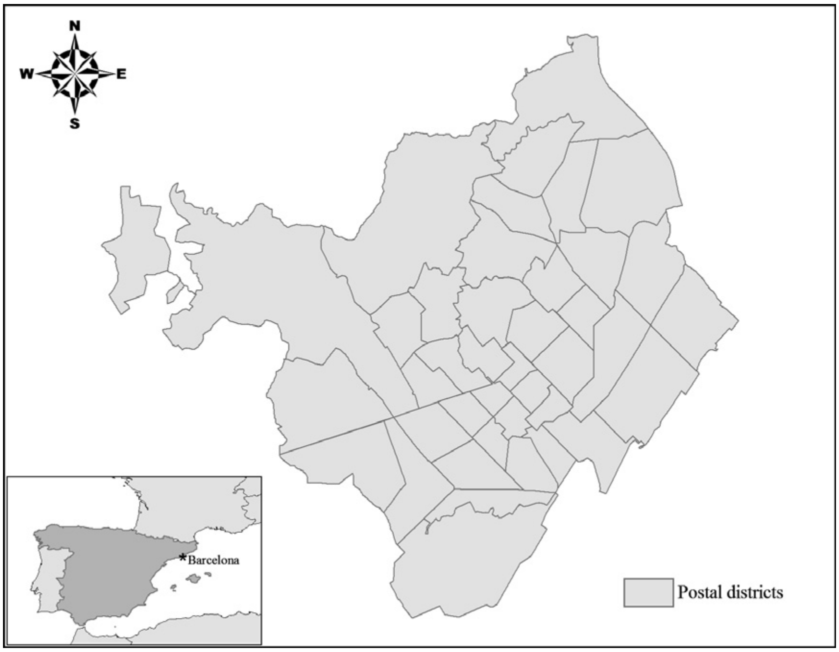

Fig. 1. City of Barcelona and its postal districts.

Table 1
Model results.

	Model 1	Model 2	Model 3	Model 4
Constant	7.6127	7.2415	7.7244	7.7758
	(0.0243)	(0.0181)	(0.0152)	(0.0594)
lsize	1.0693	1.0842	0.9473	0.9227
	(0.0053)	(0.0038)	(0.0032)	(0.0128)
lage	−0.0505	−0.0647	−0.0631	−0.0344
	(0.0010)	(0.0007)	(0.0006)	(0.0019)
D_2001		0.0853	0.0928	0.0950
		(0.0050)	(0.0040)	(0.0039)
D_2002		0.2109	0.2279	0.2307
		(0.0049)	(0.0039)	(0.0039)
D_2003		0.4004	0.4124	0.4137
		(0.0049)	(0.0039)	(0.0038)
D_2004		0.5971	0.6072	0.6085
		(0.0048)	(0.0038)	(0.0038)
District dummies	NO	NO	YES	YES
Age and size interactions with district dummies	NO	NO	NO	YES
Residuals sum squared	4365.51	2276.92	1426.97	1367.54
R^2	0.5045	0.7416	0.8380	0.8448
Adj. R^2	0.5045	0.7415	0.8379	0.8444
AIC	427294.4	394118.8	370375.6	368362.8
N	50980	50980	50980	50980

Note: Standard errors in parentheses. Models 3–4 consider the postal district (08030) with the most observations as their base category. As usual in estimates with large datasets, standard errors are small enough to make all parameters significant at 1%.

$$\text{Model 3}: \quad \ln(p) = \beta_0 + \beta_1 \ln(size) + \beta_2 \ln(age) + \sum_{t=2001}^{2004} \gamma_t T_t + \sum_{i=1}^{40} \delta_i D_i + \varepsilon \tag{7}$$

$$\text{Model4}: \quad \ln(p) = \beta_0 + \sum_{t=2001}^{2004} \gamma_t T_t + \sum_{i=1}^{40} \delta_i D_i + \sum_{i=1}^{40} \beta 1_i D_i \ln(size) + \sum_{i=1}^{40} \beta 2_i D_i \ln(age) + \varepsilon \tag{8}$$

Table 1 shows the basic results obtained with these models. The structural characteristics explain more than 50% of the model's total variance (model 1). We find that doubling the size of the dwelling involves more than doubling its price, while doubling a house's age results in a 5% fall in price. Adding time dummies (model 2) explains an additional 25% of the total variance, while incorporating spatial differentiation via the 40 postal districts (model 3) helps to explain an additional 10% of the total variance. As suggested by Bourassa et al. (2007), all estimates were performed using simple OLS regressions.[21]

Model 4 considers the spatial differentiation of the parameters (constant, age and size) for the 40 initial districts, resulting in a modest increase in adjustment compared to the previous specifications. If we examine the parameters for each district, important differences are noted (see Table 2 for a summary). Thus, the largest size parameter of a postal district is almost twice that of the smallest. The age parameter differs markedly between postal districts, with the highest being 13 times greater than the lowest. Finally, we computed Moran's I global spatial autocorrelation statistic. This was high and significant for the key parameters, indicating that these parameters present a spatial pattern. Appendix C shows the detailed results for each postal district in model 4, together with the maps of the key parameters (Figs. A3.1–A3.3).

With the results we have so far, we could proceed, as elsewhere in the literature, and attempt to combine postal districts into a small number of submarkets using Chow tests. The usual method is to use a city's administrative units, such as districts, to test whether or not submarkets matter. However, it is important to bear in mind that administrative districts are not always suitable for delineating housing submarkets: as McMillen (2010, p.139)

[21] More complex regressions could result in more efficient estimates, but as we are interested in submarket differentiation we adopt the more parsimonious option from the literature. As mentioned by an anonymous referee, incorporating district dummies may account for spatial autocorrelation, and thus diminish the utilities of using AR or SAR models.

Table 2
Distribution of municipal parameters in model 4.

	Constant	Size	Age
Min	7.358	0.616	−0.152
Q1/4	7.614	0.885	−0.081
Median	8.003	0.935	−0.070
Q3/4	8.158	0.996	−0.054
Max	9.721	1.114	−0.012
Moran's I	0.0198	0.4088	0.5661

points out, regression models for administrative units "face potential problems with omitted variables since spatial effects do not necessarily match district boundaries perfectly." If, for whatever reason, these units do not exist, then the researcher has to build them, and this is no easy task as there is an enormous amount of combinations that need to be tested. Moreover, if we were to follow a hierarchical alternative, the starting point would (as Goodman & Thibodeau, 1998, stress) condition all the results from the procedure. In this case, the HouSI heuristic appears as a methodological alternative to that of aggregating 40 postal districts into ten housing submarkets. This number allows us to compare the submarkets with the option of defining them according to the city's administrative districts.

We consider as the input for the heuristic the results from model 4, where each postal district has three parameters: a constant, and the size and age parameters. The other parameters of the heuristic are set as follows: $M = 10$, $maxitr = 5000$, and $l = 85$.

Fig. 2 shows the administrative subdivisions of the city, while Fig. 3 shows the housing submarkets resulting from HouSI, i.e. the analytical regions with the most homogeneous vectors of housing prices. A comparison of the two maps reveals that (a) analytical housing submarkets are more size-heterogeneous than the administrative districts; (b) some administrative districts can be divided in different submarkets, most notably those that lie to the east of the city (for example, district number 10); (c) large analytical housing submarkets, by contrast, can be found in the centre and to the north-west of the city. There are also major overlaps, as are to be expected of districts that have not just been built randomly in the city space. On the contrary, administrative (normative) regions are "the expression of a political will; their limits are fixed according to the tasks allocated to the territorial communities, according to the sizes of population necessary to carry out these tasks efficiently and economically, and according to historical, cultural and other factors" (Eurostat, 1999, p. 7). Specific knowledge of the city helps explain these differences, which can be classified along two main lines: those of accessibility and city transformation.

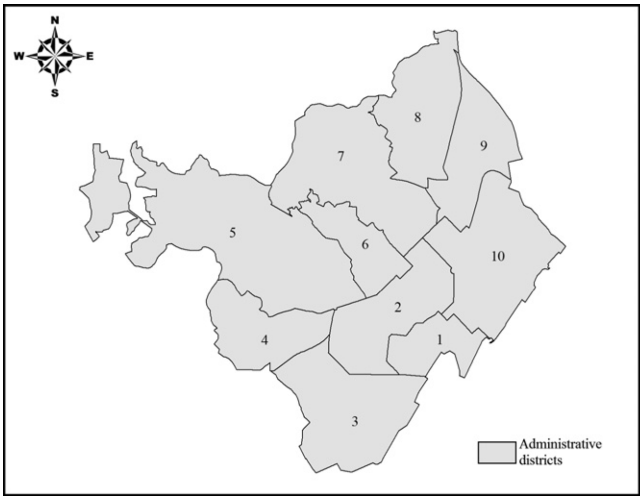

Fig. 2. Administrative districts.

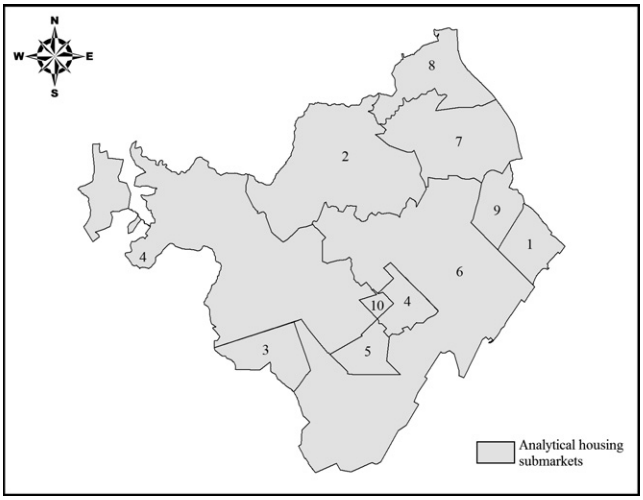

Fig. 3. Analytical housing submarkets.

Table 3
Model comparison.

	Administrative districts	HouSi solution
R^2	0.8187 (0.006)	0.8282 (0.000)
AIC	−31848.85 (0.005)	−34574,72 (0.000)
Residuals sum squared	1596.0 (0.005)	1512,9 (0.000)

Note: in parenthesis are showed the *p*-value of every statistic according to the empirical distribution computed using 1000 aggregations of the 40 original postal districts into 10 random areas.

- The city's oldest district (number 1) expanded in the nineteenth century (district number 2) and merged *vertically*, from the coastal mountain chain (to the northwest) to the sea (in the southeast), with a number of small towns that surrounded Barcelona. The city's underground system was built to reflect this growth. As a result, the housing submarkets present a monocentric structure, with regions being differentiated by their distance from the centre (see, for instance, regions 7 and 8 in Fig. 3), with property prices being highest in the city centre. This creates the two small regions for our analysis (regions 4 and 10 in Fig. 3).
- Barcelona has undergone a major urban transformation over the last 35 years (the period of analysis), at times scheduling major global events as the justification for their undertaking. This was the case of the 1992 Olympic Games and the 2004 World Forum of Cultures.[22] Urban regeneration projects have dramatically transformed the city's seafront, which had previously been a deprived, industrial area. Today, this has been replaced by region number 6 in Fig. 3, which occupies the area along the coast, and

[22] For more information on this event see www.barcelona2004.org.

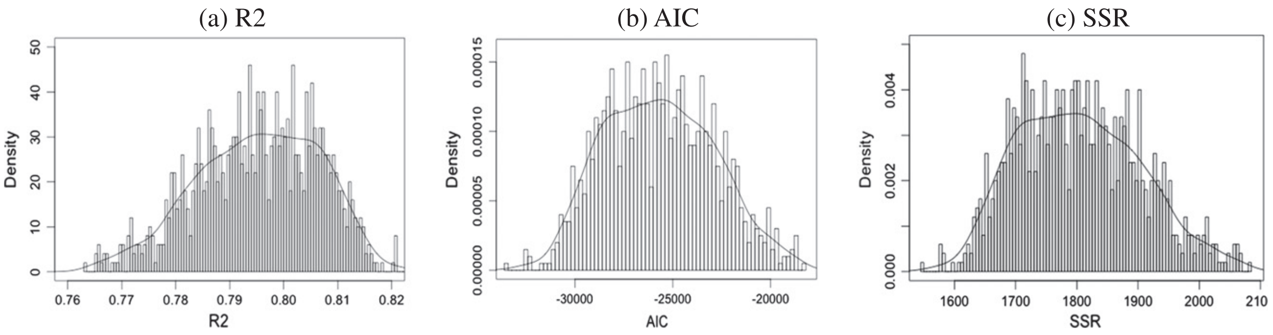

Fig. 4. Empirical distribution of R2, AIC and sum of squared residuals after 1000 random permutations.

there is a marked differentiation with the area to the east of the city (administrative district number 10 in Fig. 2), where urban regeneration has only had a partial effect.

To consider the quantitative differences between the solution provided by HouSI and the administrative division of the city, we ran a regression for each so that they could compete. We use model 4 specification (Eq. (8)), and consequently every spatial division, both the analytic submarkets derived from HouSI and the administrative units (postal districts), presents specific parameters. Table 3 shows the comparative statistics for both hedonic models: R^2; AIC statistic; and the sum of squared residuals (SSR).[23] Additionally, in order to show the goodness of our procedure, we ran 1000 random aggregations of the 40 postal districts into 10 regions in order to find the empirical distribution of each considered statistic. Table 3 shows each statistic and its empirical *p*-values in parenthesis, and Fig. 4a–c show the empirical distributions of each statistic.

Both models (administrative units and analytical regions) performed better than a single model for the entire city, thereby confirming the superiority of the submarket option. Thus, our line of argument is not so much whether or not we need to consider housing submarkets, but rather how they are best configured. We conclude that the HouSI solution provides a better result ($R^2 = 0.8282$) than that provided by the administrative division of the city ($R^2 = 0.8187$). In evaluating this 1% improvement it should be recalled that: (1) in the first set of regressions adding multiplicative dummies to all variables and districts only improved the hedonic models by 0.6%; (2) the maximum adjustment we could expect is the detailed estimate for the 40 postal district (model 4 in Table 1, $R^2 = 0.8448$); and (3) according to the empirical distributions of the statistics considered, HouSi provides a much better result than that afforded by the administrative regions.

It might be argued that these differences are not great and that, therefore, the administrative districts could just as equally serve as housing submarkets. Indeed, our findings demonstrate that such districts are not spatially random entities, but rather they are close reflections of historical movements and their social and economic realities. As such, our procedure has shown itself capable of building homogeneous housing submarkets that, at the very least, reproduce similar levels of homogeneity as those presented by these politically and socially based areas. In short, our procedure enables researchers to build efficient and theoretically consistent housing submarkets.

5. Conclusions

This paper's prime aim has been to explore new alternatives for constructing housing submarkets when merging large numbers of spatial units into a small number of such submarkets. Housing submarkets are characterised by their wide range of hedonic function parameters and, thus, two neighbourhoods can be said to belong to the same housing submarket when their hedonic functions are similar. Here, we have proposed a spatially constrained clustering heuristic, HouSI, specifically designed for identifying housing submarkets. HouSI performs parameter comparisons between polygons using a metric built as a Wald test of structural change for two subsamples of different variance plus a generalised Wald test for comparing a list of parameter vectors.

The empirical evidence indicates that HouSI serves as a potentially good alternative to the massive use of Chow tests. Future research should be directed towards endogenizing the parameter *M*, so that the optimal number of submarkets can be defined by the heuristic.

Acknowledgement

Vicente Royuela acknowledges the financial support from CICYT ECO2010-16006.

Appendix A. The generalised Wald test in a compacted form

Satorra and Neudecker (1997) develop a compact matrix expression for generalised Wald tests of equality of moment vectors in order, primarily, to evaluate the discrepancy between the weighted average of the parameters under consideration and all parameter vectors. These authors use the generalised Wald test statistic developed in Moore (1977):

$$T \equiv Nr'Q'(QYQ')^{-}Qr$$

where *N* is the total sample size, *r* is the vector of parameters of all subsamples, *Y* is the total variance covariance matrix of the parameters estimates, and finally, *Q* can be expressed as:

$$Q = \left[\begin{pmatrix} 1-\frac{N_1}{N} & \cdots & -\frac{N_G}{N} \\ \vdots & \ddots & \vdots \\ -\frac{N_1}{N} & \cdots & 1-\frac{N_G}{N} \end{pmatrix} \otimes \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}\right]$$

Satorra and Neudecker (1997) provide the following compact expression of the Moore formula:

$$T \equiv Nr'(H \otimes \overline{Y}^{-})r$$

where $\overline{Y}$ is the weighted average of the variance covariance matrix of all of all parameters vectors:

[23] We do not consider the usual common sense test (Schnare & Struyk, 1976) as we use a model in which time dummies hold for the full sample and, consequently, we do not have the standard error for each submarket. In any case, as we keep the number of variables and analytical regions constant, our statistics are in line with the expected results of the Schnare and Struyk statistic.

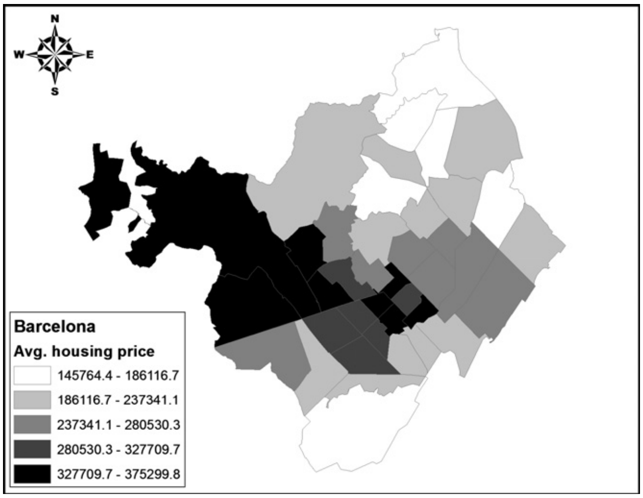

Fig. A2.1. Spatial distribution of average housing prices.

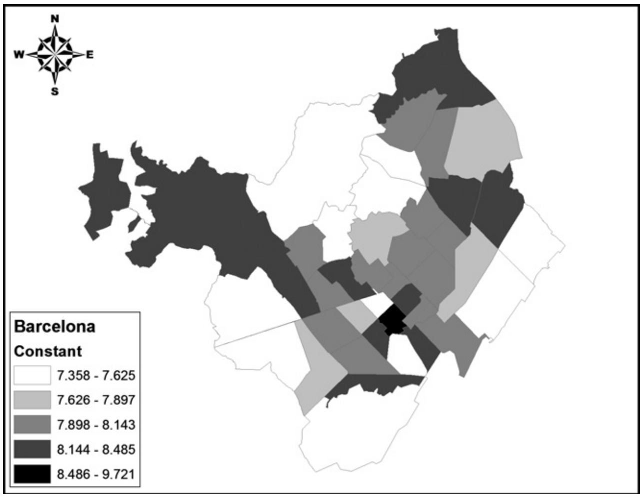

Fig. A3.1. Spatial distribution of the constant.

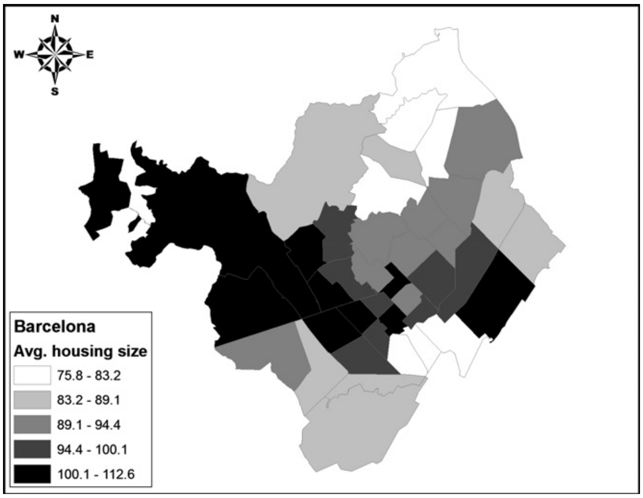

Fig. A2.2. Spatial distribution of average housing size.

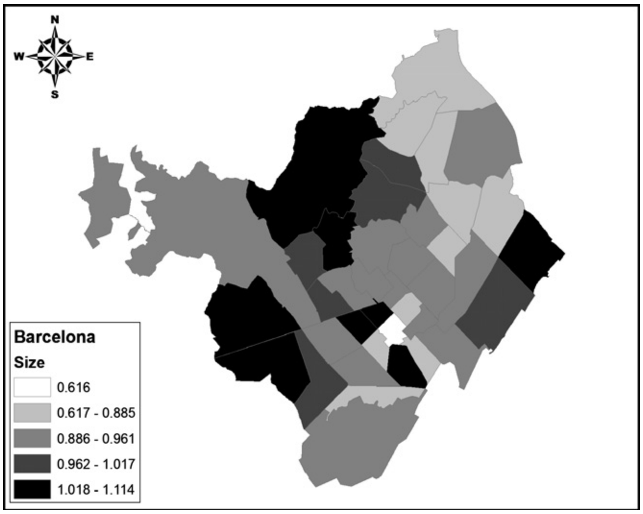

Fig. A3.2. Spatial distribution of the parameter "size".

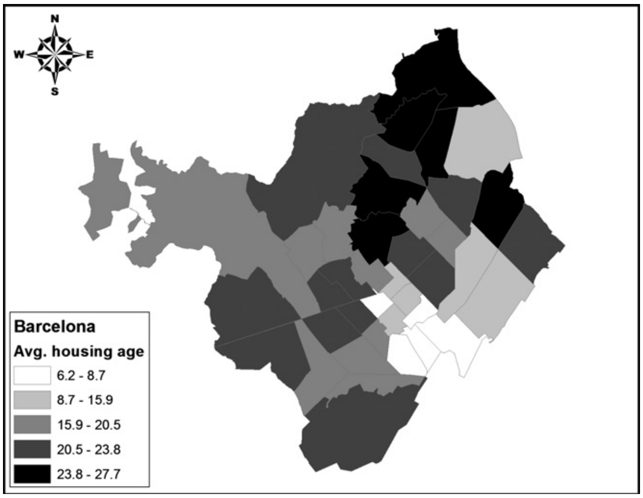

Fig. A2.3. Spatial distribution of average housing age.

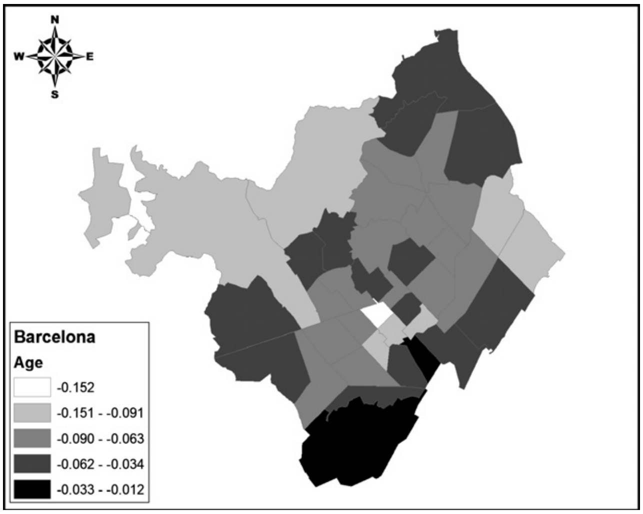

Fig. A3.3. Spatial distribution of the parameter "Age".

Table A1
Descriptive statistics.

	N	Mean	Std. dev	Min	Q1/4	Me	Q3/4	Max	Moran's I
	Complete sample								
Price	99182	238904	201823	10691	143877	195449	272779	10296041	0.4389
Size	99182	92.04	53.55	17	67	82	103	2891	0.4088
Age	99182	40.24	27.89	0	25	35	60	133	0.5661
	Restricted sample								
Price	50980	232674.3	101278	72872.7	159465.3	210502	282987	678662.9	0.5675
Size	50980	90.86	23.43	42	7	25	30	171	0.4450
Age	50980	20.74	12.88	0	74	88	104	35	0.3947

Correlations for complete sample ($N = 99182$)

	Price	Size	Age
Price	1		
Size	0.8570	1	
Age	−0.1231	−0.1153	1

Correlations for restricted sample ($n = 50980$)

	Price	Size	Age
Price	1		
Size	0.7039	1	
Age	−0.3425	−0.3040	1

$$\overline{Y} = \sum_{g=1}^{G} \frac{N_g}{N} Y_g$$

And where H can be expressed as:

$$H = \begin{bmatrix} \frac{N_1}{N} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{N_G}{N} \end{bmatrix}_{G\times G} - \begin{bmatrix} \frac{N_1}{N} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{N_G}{N} \end{bmatrix}_{G\times G} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{G\times G} \begin{bmatrix} \frac{N_1}{N} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{N_G}{N} \end{bmatrix}_{G\times G}$$

$$= \begin{bmatrix} \frac{N_1}{N} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{N_G}{N} \end{bmatrix}_{G\times G} - \begin{bmatrix} \frac{N_1}{N}\frac{N_1}{N} & \cdots & \frac{N_1}{N}\frac{N_G}{N} \\ \vdots & \ddots & \vdots \\ \frac{N_G}{N}\frac{N_1}{N} & \cdots & \frac{N_G}{N}\frac{N_G}{N} \end{bmatrix}_{G\times G} = \begin{bmatrix} \frac{N_1}{N} - \frac{N_1}{N}\frac{N_1}{N} & \cdots & -\frac{N_1}{N}\frac{N_G}{N} \\ \vdots & \ddots & \vdots \\ -\frac{N_G}{N}\frac{N_1}{N} & \cdots & \frac{N_G}{N} - \frac{N_G}{N}\frac{N_G}{N} \end{bmatrix}_{G\times G}$$

Table A3
Results by postal districts.

Postal district	Total sample		Restricted sample		Parameter estimation (Model 4)		
	Obs	Price mean	Obs	Price mean	Constant	Size	Age
8001	4939	163,936 €	691	225,964 €	7.3943	1.0359	−0.0503
8002	1442	240,335 €	194	226,604 €	8.4072	0.8099	−0.0137
8003	3176	177,262 €	466	232,473 €	7.9603	0.9264	−0.0477
8004	3638	173,406 €	1150	207,888 €	8.3032	0.8239	−0.0447
8005	3086	233,859 €	1797	276,872 €	7.5351	1.0165	−0.0552
8006	1568	377,605 €	638	327,710 €	8.2032	0.9322	−0.0738
8007	535	467,129 €	108	342,442 €	9.7207	0.6161	−0.1096
8008	476	477,396 €	154	341,962 €	7.5253	1.1138	−0.1516
8009	618	368,425 €	169	316,995 €	8.4851	0.8455	−0.0554
8010	897	380,299 €	334	336,836 €	8.1326	0.9447	−0.0916
8011	1317	297,659 €	419	290,822 €	8.3944	0.8840	−0.1043
8012	2448	218,021 €	552	249,191 €	7.9583	0.9399	−0.0592
8013	2963	247,257 €	1501	269,886 €	8.0932	0.9246	−0.0791
8014	3775	207,871 €	2012	234,320 €	7.7486	0.9748	−0.0648
8015	3204	249,352 €	1190	286,759 €	8.1152	0.9236	−0.0787
8016	3123	172,156 €	2182	177,936 €	8.1216	0.8772	−0.0852
8017	1694	585,754 €	565	370,672 €	8.2131	0.9490	−0.0957
8018	2713	232,390 €	1969	253,287 €	7.8971	0.9330	−0.0660
8019	2351	194,986 €	1396	212,538 €	7.3575	1.0675	−0.1170
8020	3982	180,216 €	2963	181,830 €	8.2392	0.8499	−0.0916
8021	1309	590,552 €	377	374,267 €	8.0040	1.0007	−0.0840
8022	1412	480,913 €	506	346,327 €	7.9781	0.9779	−0.0510
8023	1905	303,605 €	1077	280,530 €	7.5579	1.0318	−0.0582
8024	2932	224,661 €	1511	237,341 €	7.8762	0.9565	−0.0735
8025	3529	229,670 €	1699	249,062 €	8.0917	0.9098	−0.0619
8026	3095	220,589 €	1773	243,717 €	8.1425	0.8849	−0.0635
8027	3651	207,275 €	2266	218,683 €	8.2291	0.8618	−0.0737
8028	4378	232,714 €	2824	255,911 €	7.4459	1.0430	−0.0469
8029	2844	280,890 €	1544	307,553 €	8.0238	0.9613	−0.0798
8030	4960	200,215 €	3978	204,568 €	7.7758	0.9227	−0.0344
8031	3311	189,312 €	1948	209,861 €	7.5823	0.9944	−0.0726
8032	4074	183,048 €	2671	182,142 €	7.5423	0.9928	−0.0675
8033	2372	139,238 €	1576	145,764 €	8.2710	0.7663	−0.0414
8034	1370	569,306 €	580	375,300 €	7.5141	1.0877	−0.0475
8035	1418	223,493 €	898	201,949 €	7.4239	1.0428	−0.0911
8036	1260	313,755 €	397	324,886 €	7.7135	1.0295	−0.0799
8037	834	350,395 €	284	351,656 €	8.0816	0.9373	−0.0758
8038	1845	187,176 €	1453	186,117 €	7.6248	0.9253	−0.0117
8041	1774	207,705 €	1152	218,857 €	8.0093	0.9171	−0.0728
8042	2964	153,789 €	2016	162,767 €	8.0022	0.8725	−0.0592
Total	99182	238,904 €	50980	232,674 €	–	–	–

Appendix B. Descriptive statistics

See Figs. A2.1–A2.3 and Table A1.

Appendix C. Results by postal districts

See Figs. A3.1–A3.3 and Table A3.

References

Adair, A. S., Berry, J. N., & McGreal, W. S. (1996). HedonicModeling, housing submarkets and residential valuation. *Journal of Property Research, 13*(1), 67–83.

Altman, M. (1997). Is automation the answer: The computational complexity of automated redistricting. *Rutgers Computer and Law Technology Journal, 23*(1), 81–142.

Bergey, P., Ragsdale, C., & Hoskote, M. (2003). A decision support system for the electrical power districting problem. *Decision Support System, 36*, 1–17.

Blais, M., Lapierre, S., & Laporte, G. (2003). Solving a home-care districting problem in an urban setting. *Journal of the Operational Research Society, 54*(11), 1141–1147.

Bond, S. A., & Hwang, S. S. (2007). Smoothing, nonsynchronous appraisal and cross-sectional aggregation in real estate price indices. *Real Estate Economics, 35*(3), 349–382.

Bourassa, S. C., Cantoni, E., & Hoesli, M. (2007). Spatial dependence, housing submarkets, and house prices. *Journal of Real Estate Finance and Economics, 35*(2), 143–160.

Bourassa, S. C., Hamelink, F., Hoesli, M., & MacGregor, B. D. (1999). Defining housing submarkets. *Journal of Housing Economics, 8*(2), 160–183.

Bozkaya, B., Erkut, E., & Laporte, G. (2003). A Tabu Search heuristic and adaptive memory procedure for political districting. *European Journal of Operational Research, 144*, 12–26.

Byfuglien, J., & Nordgard, A. (1973). Region-building: A comparison of methods. *Norwegian Journal of Geography, 27*, 127–151.

Caro, F., Shirabe, T., Guignard, M., & Weintraub, A. (2004). School redistricting: Embedding GIS tools with integer programming. *Journal of the Operational Research Society, 55*(8), 836–849.

Cheshire, P., & Sheppard, S. (1995). On the price of land and the value of amenities. *Economica, 62*, 247–267.

Cliff, A. D., & Hagget, P. (1970). On the efficiency of alternative aggregations in region-building problems. *Environment and Planning, 2*, 285–294.

Cliff, A. D., Haggett, P., Ord, J. K., Bassett, K. A., & Davies, R. B. (1975). *Elements of spatial structure: A quantitative approach*. New York: Cambridge University Press.

Comber, A., Carver, S., Fritz, S., McMorran, R., Washtell, J., & Fisher, P. (2010). Different methods, different wilds: Evaluating alternative mappings of wildness using fuzzy MCE and Dempster-Shafer MCE. *Computers Environment and Urban Systems, 34*(2), 142–152.

Dietrich, J. R., Harris, M. S., & MullerIII, K. A. (2000). The reliability of investment property fair value estimates. *Journal of Accounting and Economics, 30*, 125–158.

Duque, J. C., Anselin, L., & Rey, S. J. (2011). The max-p-regions problem. *Journal of Regional Science*. http://dx.doi.org/10.1111/j.1467-9787.2011.00743.x.

Duque, J. C., Artis, M., & Ramos, R. (2006). The ecological fallacy in a time series context: Evidence from Spanish regional unemployment rates. *Journal of Geographical Systems, 8*, 391–410.

Duque, J. C., Church, R. L., & Middleton, R. S. (2011). The p-regions problem. *Geographical Analisis, 43*(1), 104–126.

Duque, J. C., Ramos, R., & Surinach, J. (2007). Supervised regionalization methods: A survey. *International Regional Science Review, 30*, 195–220.

Eurostat (1999). *Regio database, user's guide, methods and nomenclature*. Luxembourg: Official Publication Office.

Fan, G. Z., Ong, S. E., & Koh, H. C. (2006). Determinants of house price: A decision tree approach. *Urban Studies, 43*(12), 2301–2316.

Ferligoj, A., & Batagelj, V. (1982). Clustering with relational constraint. *Psychometrika, 47*(4), 413–426.

Fischer, M. M. (1980). Regional taxonomy: A comparison of some hierarchic and non-hierarchic strategies. *Regional Science and Urban Economics, 10*, 503–537.

Glover, F. (1977). Heuristic for integer programming using surrogate constraints. *Decision Science, 8*, 156–166.

Glover, F. (1989). Tabu search. Part I. ORSA. *Journal on Computing, 1*, 190–206.

Glover, F. (1990). Tabu search. Part II. ORSA. *Journal on Computing, 2*, 4–32.

Goodman, A. C., & Kawai, M. (1982). Permanent income, hedonic prices, and demand for housing: New evidence. *Journal of Urban Economics, 12*(2), 214–237.

Goodman, A. C., & Thibodeau, T. T. (1998). Housing market segmentation. *Journal of Housing Economics, 7*, 121–143.

Goodman, A. C., & Thibodeau, T. T. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics, 12*, 181–201.

Goodman, A. C., & Thibodeau, T. T. (2007). The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics, 35*(2), 209–232.

Gordon, A. D. (1996). A survey of constrained classification. *Computational Statistics & Data Analysis, 21*, 17–29.

Gordon, A. D. (1999). *Classification* (2nd ed). Boca Raton, FL: Chapman & Hall-CRC.

Gravel, N., Michelangeli, A., & Trannoy, A. (2006). Measuring the social value of local public goods: An empirical analysis within Paris metropolitan area. *Applied Economics, 38*(16), 1945–1961.

Gunn, S. C. (2007). Green belts: A review of the regions' responses to a changing housing Agenda. *Journal of Environmental Planning and Management, 50*(5), 595–616.

Hand, M. S., Thacher, J. A., & McCollum, D. W. (2008). Intra-regional amenities, wages, and home prices: The role of forests in the southwest. *Land Economics, 84*(4), 635–651.

Hansen, P., Jaumard, B., Meyer, C., Simeone, B., & Doring, V. (2003). Maximum split clustering under connectivity constraints. *Journal of Classification, 20*, 143–180.

Hansz, J. A., & Diaz, J. III, (2003). Valuation bias in commercial appraisal: A transaction price feedback experiment. *Real Estate Economics, 29*, 553–565.

Horn, M. (1995). Solution techniques for large regional partitioning problems. *Geographical Analysis, 27*(3), 230–248.

Hwang, S., & Thill, J. C. (2009). Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning B – Planning & Design, 36*(5), 865–882.

Islam, K. S., & Asami, Y. (2009). Housing market segmentation: A review. *Review of Urban & Regional Development Studies, 21*(2–3), 93–109.

Jenkins, W. (1978). *Policy analysis: A political and organizational perspective*. London: Martin Robertson.

Johnston, R. J. (1968). Choice in classification: The subjectivity of objective methods. *Annals of the AAG, 58*, 575–589.

Jones, C., Leishman, C., & Watkins, C. (2003). Structural change in a local urban housing market. *Environment and Planning A, 35*(7), 1315–1326.

Jones, C., Leishman, C., & Watkins, C. (2004). Intra-urban migration and housing submarkets: Theory and evidence. *Housing Studies, 19*(2), 269–283.

Kauko, T. (2004). A comparative perspective on urban spatial housing market structure: Some more evidence of local sub-markets based on a neural network classification of Amsterdam. *Urban Studies, 41*(13), 2555–2579.

Keane, M. (1975). The size of region-building problem. *Environment and Planning A, 7*, 575–577.

Kirkpatrick, S., Gelatt, C., & Vecchi, M. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

Lankford, P. (1969). Regionalization: Theory and alternative algorithms. *Geographical Analysis, 1*, 196–212.

Lefkovitch, L. (1980). Conditional clustering. *Biometrics, 36*(1), 43–58.

Legendre, P. (1987). *Developments in numerical ecology. NATO ASI Series. Constrained clustering* (Vol. G 14, pp. 289–307). Berlin: Springer (chap.).

Li, X. (2007*). P-region based estimation of disease rates*. Master Thesis. San Diego State University.

Lin, L., & Gen, G. (2009). Auto-tuning strategy for evolutionary algorithms: Balancing between exploration and exploitation. *Soft Computing, 13*(2), 157–168.

Maravalle, M., & Simeone, B. (1995). A spanning tree heuristic for regional clustering. *Communications in Statistics – Theory and Methods, 24*, 625–639.

Mashinchi, M. H., Orgun, M. A., & Pedrycz, W. (2011). Hybrid optimization with improved tabu search. *Applied Soft Computing, 11*(2), 1993–2006.

McMillen, D. P. (2010). Issues in spatial data analysis. *Journal of Regional Science, 50–1*, 119–141.

Moore, D. S. (1977). Generalized inverses, Walds method, and construction of chi-squared tests of fit. *Journal of the American Statistical Association, 72*, 131–137.

Murtagh, F. (1985). A survey of algorithms for contiguity-constrained clustering and related problems. *Computer Journal, 28*, 82–88.

Murtagh, F. (1992). Contiguity-constrained clustering for image analysis. *Pattern Recognition Letters, 13*, 677–683.

Nagel, S. (1965). Simplified bipartisan computer redistricting. *Stanford Law Review, 17*(5), 863–899.

Ohtani, K., & Toyoda, T. (1985). Small sample properties of tests of equality between sets of coefficients in 2 linear regressions under heteroscedasticity. *International Economic Review, 26*, 37–44.

Openshaw, S. (1977). A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modeling. *Transactions of the Institute of British Geographers, 2*, 459–472.

Openshaw, S., & Rao, L. (1995). Algorithms for reengineering 1991 census geography. *Environment and Planning A, 27*, 425–446.

Páez, A. (2009). Recent research in spatial real estate hedonic analysis. *Journal of Geographical systems, 11*, 311–316.

Pavlov, A. D. (2000). Space-varying regression coefficients: A semi-parametric approach applied to real estate markets. *Real Estate Economics, 28*(2), 249–283.

Pezzella, F., Bonanno, R., & Nicoletti, B. (1981). A system approach to the optimal health-care districting. *European Journal of Operational Research, 8*, 139–146.

Redfearn, C. L. (2009). How informative are average effects? Hedonic regression and amenity capitalization in complex urban housing markets. *Regional Science and Urban Economics, 39*, 297–306.

Ricca, F., & Simeone, B. (2008). Local search algorithms for political districting. *European Journal of Operational Research, 189*, 1409–1426.

Ríos-Mercado, R., & Fernández, E. (2009). A reactive GRASP for a commercial territory design problem with multiple balancing requirements. *Computers & Operations Research, 36*, 755–776.

Royuela, V., & Vargas, M. (2009). Defining housing market areas using commuting and migration algorithms. Catalonia (Spain) as an applied case study. *Urban Studies, 46*(11), 2381–2398.

Sammons, R. (1978). *A simplistic approach to the redistricting problem. Spatial representation and spatial interaction* (pp. 71–94). Leiden: M. Nijho Social Sciences Division (chap.).

Satorra, A., & Neudecker, H. (1997). Compact matrix expressions for generalized Wald tests of equality of moment vectors. *Journal of Multivariate Analysis, 63*, 259–276.

Schnare, A., & Struyk, R. (1976). Segmentation in urban housing markets. *Journal of Urban Economics, 3*, 146–166.

Sunding, D. L., & Swoboda, A. M. (2010). Hedonic analysis with locally weighted regression: An application to the shadow cost of housing regulation in Southern California. *Regional Science and Urban Economics, 40*, 550–573.

Taylor, P. J. (1973). Some implications of spatial organization of elections. *Transactions of the Institute of British Geographers, 60*, 121–136.

Tobler, W. R. (1970). Computer movie simulating urban growth in Detroit region. *Economic Geography, 46*, 234–240.

Toyoda, T., & Ohtani, K. (1986). Testing equality between sets of coefficients after a preliminary test for equality of disturbance variances in 2 linear regressions. *Journal of Econometrics, 31*, 67–80.

Tu, Y., Sun, H., & Yu, S. M. (2007). Spatial autocorrelations and urban housing market segmentation. *Journal of Real Estate Finance and Economics, 34*(3), 385–406.

Watkins, C. (2001). The definition and identification of housing submarkets. *Environment and Planning A, 33*, 2235–2253.

Weeks, J. R., Hill, A. G., Getis, A., & Stow, D. (2006). Ethnic residential patterns as predictors of intra-urban child mortality inequality in Accra, Ghana. *Urban Geography, 27*(6), 526–548.

Weeks, J. R., Hill, A. G., Stow, D., Getis, A., & Fugate, D. (2007). Can we spot a neighborhood from the air? Defining neighborhood structure in Accra, Ghana". *GeoJournal, 69*(1–2), 9–22.

Weise, T. (2009). *Global optimization algorithms – Theory and application* (2nd ed.). Germany: University of Kassel, Distributed Systems Group, Self-Published.

Wilkinson, R. K. (1973). House prices and the measurement of externalities. *Economic Journal, 83*, 72–86.

Williams, J. (1995). Political redistricting: A review. *Papers in Regional Science, 74*(1), 13–40.

Wise, S., Haining, R., Ma, J. (1997). Recent developments in spatial analysis: Spatial statistics, behavioural modelling, and computational intelligence. In Manfred M. Fischer & Arthur Getis (Eds.), *Regionalisation tools for exploratory spatial analysis of health data* (pp. 83–100). New York: Springer (chap.).

Yamada, T. (2009). A mini-max spanning forest approach to the political districting problem. *International Journal of Systems Science, 40*(5), 471–477.

Yiu, C. Y., & Wong, S. K. (2005). The effects of expected transport improvements on housing prices. *Urban Studies, 42*(1), 113–125.

Zoltners, A., & Sinha, P. (1983). Sales territory alignment: A review and model. *Management Science, 29*(11), 1237–1256.