Endogenous Spatial Regimes

Luc Anselin* and Pedro Amaral†

Abstract

In spatial econometrics, the treatment of spatial heterogeneity can be approached from a continuous or a discrete perspective. In a continuous approach, represented by methods such as geographically weighted regression (GWR) and Bayesian varying coefficient specifications, the model coefficients are allowed to vary smoothly over space. In contrast, in a discrete perspective, referred to as *spatial regimes*, the coefficients vary by discrete subregions of the data.

Whereas the estimation of spatial regime regressions is well understood, the delineation of the regimes themselves remains a topic of active interest. Broadly speaking, three approaches can be distinguished: exogenous regimes, determined a priori (e.g., administrative regions); data-driven regimes, obtained as the result of a clustering exercise; and endogenous regimes, where the coefficients and the regime allocation are obtained jointly, e.g., as the result of a finite mixture regression. One drawback of most data-driven and endogenous regime delineation is that the results do not necessarily satisfy a spatial contiguity constraint, i.e., observations are grouped together that are not spatially connected.

In this paper, we propose a new heuristic to determine the spatial regimes endogenously, as an extension of the well-known SKATER algorithm for spatially constrained clustering. This guarantees that the resulting regimes consist of contiguous observations. We outline the method and apply it in the context of the determination of housing submarkets, which is represented by a rich literature in applied spatial econometrics. We use a well-known Kaggle data set as the empirical example, which contains observations on house sales in King County, Washington. We compare the estimation of a hedonic house price model using the new endogenous spatial regimes approach to a range of more traditional methods, including pooled regression, the use of administrative districts, data-driven regimes based on a-spatial and spatial clustering of explanatory variables, and finite mixture regression. We evaluate the results in terms of fit and assess the trade-offs between the spatial and a-spatial approaches.

Keywords: spatial heterogeneity, spatial regimes, spatially constrained clustering, SKATER, housing submarkets

*Center for Spatial Data Science, University of Chicago, Chicago IL – anselin@uchicago.edu

†Department of Economics, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil – pedroamaral@cedeplar.ufmg.br

Endogenous Spatial Regimes

1 Introduction

As is well known, in spatial econometrics, the two fundamental spatial effects that require a specialized methodology are spatial dependence and spatial heterogeneity (Anselin, 1988). Of the two, spatial dependence has arguably led to the bulk of the results in the literature, although spatial heterogeneity has received considerable attention as well, especially since the renewed focus on the local in quantitative geography (e.g., Fotheringham, 1997; Lloyd, 2010, among others).

Fundamentally, a distinction can be made between a discrete and a continuous perspective on spatial heterogeneity. In the continuous approach, regression model coefficients vary smoothly over the locations (observations), either in a deterministic or in a random fashion. An early approach to continuous spatial heterogeneity was the spatial expansion method outlined in the work of Casetti (Casetti, 1972, 1997), where regression coefficients are expressed as a function of other variables, somewhat similar to the specification of a hierarchical linear model (Raudenbush and Bryk, 2002). A slightly different approach is based on the idea behind local regression (Cleveland and Devlin, 1988; Loader, 1999), in which local estimates are based on a subset of the observations. Whereas in the original local regression this subset is defined over one of the covariates, in the geographically weighted regression (GWR) proposed by Fotheringham et al. (1998, 2002), the coefficient estimate for each location is based on covariates in geographically neighboring observations, using a kernel weighting approach. Randomly varying coefficients can be found primarily in the Bayesian literature (e.g., the review in Gelfand et al., 2003). A recent comparison of GWR and Bayesian spatially varying coefficient models is given in Wolf et al. (2018).

In the current paper, we focus on discrete heterogeneity, in the form of *spatial regimes*, as defined in Anselin (1988) (for a more recent review, see Anselin and Rey, 2014, Chapters 12 and 13). Using the notation from Anselin and Rey (2014), the basic spatial regimes specification can be expressed

as:[1]

$$y_{ij} = \alpha_j + \mathbf{x'}_{ij}\beta_j + \epsilon_{ij},$$

for $i = 1, \ldots, n$, with n as the total number of observations, $\mathbf{x'}$ a vector containing the explanatory variables and ϵ as the error term. The observations are also indexed by j, the regime to which each observation belongs, with $j = 1, \ldots, J$, where there are a total of J different spatial regimes. Each regime consists of a subset of observations and each observation belongs to one and only one regime (there are no empty regimes). *Spatial* regimes are made up of spatially contiguous observations or *regions*, but in general, regimes could be any stratification of the observations into subcategories.

The implication of this specification is that each *regime* j has its own intercept α_j and set of slope coefficients β_j. The standard assumption is homoskedasticity, where $\text{Var}[\epsilon_{ij}] = \sigma^2$. However, it is typically more realistic to allow for groupwise heteroskedasticity, such that $\text{Var}[\epsilon_{ij}] = \sigma_j^2$ for each j. Alternatively, in the most general case, a fully heteroskedastic error may be assumed, such that $\text{Var}[\epsilon] = \mathbf{\Sigma}$.

In the simplest specification, only the intercept varies between regimes, which is sometimes referred to as cross-sectional spatial fixed effects (Kuminoff et al., 2010; Anselin and Arribas-Bel, 2013), as distinct from spatial fixed effects in a panel data setting (Lee and Yu, 2011; Elhorst, 2014).

In essence, when groupwise heteroskedasticity is assumed, the specification of spatial regimes is equivalent to a different regression for each group. This is the standard approach to test for structural stability in a regression model (e.g., Chow, 1960). The *spatial* regimes setup is only special in the sense that the definition of the different regimes is based on spatial structure. In most other respects, it is equivalent to the treatment of structural instability in standard (non-spatial) regression analysis. An additional complication is that often spatial heterogeneity occurs jointly with spatial dependence, which complicates specification testing and estimation (Anselin, 1988).

Different slope coefficients between regimes suggest that the response of the dependent variable to the explanatory variables is not homogeneous. This heterogeneity could be due to several factors,

[1] To keep the notation simple, the same linear functional form is assumed for all observations, but the coefficients are allowed to vary.

but in the spatial regimes model only the presence of spatial structural instability is indicated, not the reasons why it occurs. A simple test for such heterogeneity in spatial econometric specifications is a spatialized version of the Chow test (Anselin, 1990).

In applied spatial econometrics, the most common application of spatial heterogeneity is in the consideration of submarkets in the estimation of hedonic house price models. Such market segmentation may result from inelasticities in both supply and demand, resulting in spatially varying marginal prices, i.e., different coefficients in the hedonic price regression model.

The literature on the econometric treatment of housing submarkets is vast, but a comprehensive review is beyond the current scope. Classic references that emphasize the spatial aspects include Goodman and Thibodeau (1998, 2003, 2007), and Bourassa et al. (1999, 2003, 2007, 2010). Extensive literature reviews can be found in Anselin and Lozano-Gracia (2009), Helbich et al. (2013), and Bhattacharjee et al. (2016), among others.

The typical econometric treatment of spatial regimes is carried out in two stages. In the first, the regimes are delineated, either based on some exogenous classification (e.g., administrative areal units) or derived from the data (using various clustering techniques). In the second stage, the heterogeneity of the regimes is taken into account in the model estimation, in the form of different intercept and/or slope coefficients. If considered, the treatment of spatial effects is implemented in the second stage.

In the first stage, the resulting regimes often do not yield solutions where observations in each regime are also spatially contiguous, especially when a standard clustering method is applied, such as K-means clustering. Whether or not this is desired depends on the context, although there is no consensus. For example, in the housing submarket literature, the advantages of spatially delineated submarkets are touted by some (e.g., Bourassa et al., 2003), whereas others see it rather as an unnecessary constraint (e.g., Belasco et al., 2012). In this paper, we focus on regimes that enforce the contiguity constraint, i.e., *spatial* regimes.

Alternatives to the two-stage approach consist of methods where the regime determination and the coefficient estimation are carried out jointly, which we term *endogenous* regimes. An early example in the housing literature is the application of finite mixture models, e.g., by Ugarte et al.

(2004) and Belasco et al. (2012). However, in these applications, the spatial constraints are not satisfied.

In this paper, we present a new approach to let the data determine the spatial regimes endogenously by means of the integration of the regression fit optimization into a spatially constrained clustering algorithm. Specifically, we propose a heuristic in which the regression goodness-of-fit is used in the objective function of the SKATER algorithm (Spatial 'K'luster Analysis by Tree Edge Removal) of Assunção et al. (2002, 2006).

In the remainder of the paper, we first review a number of approaches to delineate spatial regimes, with selected examples pertaining to housing market segmentation. We categorize these as exogenous regimes, data-driven regimes and endogenous regimes. This is followed by a technical discussion of estimation methods, with particular attention to finite mixture models and our new spatially constrained endogenous regimes. These methods are compared in an empirical application of hedonic house price regression using the well-known Kaggle data set with house sale prices and characteristics for King County, Washington. We close with some concluding remarks.

2 The Delineation of Spatial Regimes

2.1 Exogenous Regimes

The simplest delineation of spatial regimes is when the grouping of observations is taken to be exogenous, determined a priori, based on some criteria that are totally outside the subject of the study. Typically, these are administrative areas, such as census tracts or neighborhood definitions, or clearly defined subregions, such as Baltimore City vs Baltimore County, or eastern, southern and western states in the U.S., as in the examples in Anselin and Rey (2014, Chapter 12). Early applications in the housing submarket literatures are the use of administrative boundaries in Bourassa et al. (1999), zip code zones and school districts in Goodman and Thibodeau (2003), and counties in Brasington and Hite (2005), among many applications.

The typical approach taken in empirical practice is to deal with these subareas by means of spatial fixed effects, i.e., only the intercept is allowed to vary, or a dummy variable quantifies the

difference with a regime selected as the base case.

2.2 Data-Driven Regimes

We refer to data-driven regimes as those approaches where the regional subdivisions are the result of a clustering exercise, such as K-means, typically applied to the explanatory variables in the model (sometimes including the dependent variable as well). In the literature on hedonic house price models this includes housing characteristics or characteristics of the owners/sellers (if available).

Sometimes, the clustering is applied to larger spatial units that encompass the individual house observations. For example, census tract characteristics can be used to obtain a classification of tracts into "neighborhoods", which then form the basis for the regimes applied to the individual observations (e.g., house sales within the tracts).

When many variables are available, dimension reduction is typically carried out and the clustering is applied to a subset of the principal components of the original variables. An early example of this approach is contained in Bourassa et al. (2003). However, the resulting clusters typically do not form proper spatial areal units. Such a spatially contiguous solution may be facilitated by including the x-y coordinates of the observations among the cluster variables, as in Bourassa et al. (2010), although this still does not offer a guarantee.

In order to obtain clusters where the included observations are spatially contiguous, one of the several spatially constrained clustering methods can be applied, such as SKATER (Assunção et al., 2002, 2006), Redcap (Guo, 2008; Guo and Wang, 2011), AZP (Openshaw, 1977; Openshaw and Rao, 1995), or max-p (Duque et al., 2012). However, to date, such applications are very rare in the housing submarket literature. An exception is Helbich et al. (2013), where the SKATER algorithm is applied to principal components of coefficient surfaces obtained from geographically weighted regression to derive the submarket definitions. Once the regimes are delineated, the heterogeneity can be accounted for in the hedonic regression.

A slightly different approach is taken by Bhattacharjee et al. (2016). Their focus is on the delineation of housing submarkets as such. To that effect, they develop a new spatial functional

observation is allocated to the regime for which its posterior probability is the largest. However, this allocation does not ensure spatial contiguity. Standard regime regression can then be applied to this hard assignment. Technical details are considered in Section 3.2.

The heuristic we propose also jointly carries out the allocation to regimes and the estimation of the regression coefficients. It is similar in spirit to the neighborhood analysis strategy recently outlined in Olson et al. (2021), which integrates the regression estimation into an integer programming framework, following the classification and regression via integer optimization (CRIO) methodology from the machine learning literature (Bertsimas and Shioda, 2007). However, this approach does not enforce contiguity, thus resulting in spatially disparate cluster allocations. Instead, we integrate the regression fit into the objective function of the SKATER spatially constrained clustering algorithm, thereby ensuring the construction of proper spatial regimes. We provide a technical discussion in Section 3.3.

3 Regime Regression

3.1 The Textbook Case

The elements of spatial regime regression estimation are fairly standard, but for the sake of clarity, we briefly outline the basics in what follows.[3] In our exposition, we will use a simplified specification where the structural instability pertains to two subregions, $j = 1, 2$. We can express the two regressions in pooled form as:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & 0 \\ 0 & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix},$$

where $\mathbf{y_1}$ and $\mathbf{y_2}$ are vectors of observations on the dependent variable, respectively of dimensions $n_1 \times 1$ and $n_2 \times 1$ (with $n_1 + n_2 = n$), $\mathbf{X_1}$ and $\mathbf{X_2}$ are $n_1 \times k$ and $n_2 \times k$ matrices of observations on the k explanatory variables, β_1 and β_2 are $k \times 1$ vectors of the regression coefficients in each subset

[3] This section is based on Chapters 12 and 13 in Anselin and Rey (2014), to which we refer for further technical details.

(including a constant term), and ϵ_1 and ϵ_2 are $n_1 \times 1$ and $n_2 \times 1$ vectors of error terms.

With groupwise heteroskedasticity, i.e., a separate error variance for each regime, estimation is equivalent to a separate regression in each regime. A fully general form of heteroskedasticity can be introduced as well, resulting in an application of feasible generalized least squares (FGLS). A test for the constancy of the regression coefficients, either individually or jointly, can be based on the classic Chow test (Chow, 1960).

Spatial dependence can be introduced into this specification in two main ways. In the first, the spatial autoregressive coefficient is fixed throughout the sample. For example, in the two-regime case, a spatial lag specification would be:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \rho \mathbf{W} \begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} + \begin{bmatrix} \mathbf{X_1} & 0 \\ 0 & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix},$$

with a fixed spatial autoregressive coeffcient ρ and associated spatially lagged dependent variable $\mathbf{Wy}$.

This specification is appropriate when it is assumed that a single spatial process operates on the complete data set. Consequently, there is only one spatial autoregressive coefficient ρ, even though the other (non-spatial) parameters in the model may vary across regimes.

Extending the same rationale to the spatial autoregressive error model, the regime specification for the non-spatial parameters boils down to the familiar expression:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \mathbf{X_1} & 0 \\ 0 & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix},$$

where the error terms follow a spatial autoregressive process, as in:

$$\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} = \lambda \mathbf{W} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix}.$$

with $\mathbf{u}_{1,2}$ as the idiosyncratic error vectors belonging to each regime. Again, the underlying as-

sumption is that a unique spatial process drives the dependence across the full data set, determined by a single autoregressive parameter λ with an associated spatial weights matrix $\mathbf{W}$ applied to the stacked error terms $\boldsymbol{\epsilon_1}$ and $\boldsymbol{\epsilon_2}$.

A different perspective is offered when the spatial parameters are allowed to vary across regimes. For the spatial lag model, the specification in our two-region example then becomes:

$$\begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} = \begin{bmatrix} \rho_1 \mathbf{W_1} & 0 \\ 0 & \rho_2 \mathbf{W_2} \end{bmatrix} \begin{bmatrix} \mathbf{y_1} \\ \mathbf{y_2} \end{bmatrix} + \begin{bmatrix} \mathbf{X_1} & 0 \\ 0 & \mathbf{X_2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix},$$

with the spatial weights $\mathbf{W}_{1,2}$ corresponding to the separate weights for each regime, pertaining only to those observations that belong to the regime.

For the spatial autoregressive error process, the counterpart is:

$$\begin{bmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \end{bmatrix} = \begin{bmatrix} \lambda_1 \mathbf{W_1} & 0 \\ 0 & \lambda_2 \mathbf{W_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{\epsilon_1} \\ \boldsymbol{\epsilon_2} \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \end{bmatrix},$$

where again the weights are the regime weights, with matching autoregressive coefficients.

The interpretation of spatial parameters that vary by regime is somewhat more complex than for the regression parameters. It is important that the assumption of separate spatial processes driving the regimes is realistic. This implies that there are no spill-overs between regimes and any dependencies are fully contained within each regime. In practice, it is usually more appropriate to assume a single spatial process across all regimes.

Estimation of the spatial models follows the usual principles, either employing Maximum Likelihood or Generalized Method of Moments. Details can be found in Anselin and Rey (2014), among others. A test on the hypothesis of constant coefficients across regimes can be based on the spatial version of the Chow test (Anselin, 1990).

3.2 Finite Mixture Models

The standard finite mixture model in a regression context formulates the conditional probability of a given observation for the dependent variable y as a weighted sum of conditional probabilities for each regime j, weighted by a probability π_j. More formally, this can be expressed as:

$$h(y|x,\theta) = \sum_{j=1}^{J} \pi_j f(y|x,\beta_j,\sigma_j^2),$$

with $\pi_j \geq 0$ and $\sum_{j=1}^{J} \pi_j = 1$, and h and f as conditional probability densities. The latter is typically assumed to be Gaussian in a linear regression context. Each regime has its own set of regression coefficients β_j and error variance σ_j^2. For ease of notation, all the parameters (i.e., the π_j, β_j and σ_j^2 for each regime) are grouped in the vector θ.[4]

The posterior probability that an observation i belongs to regime j is:

$$P(i \in j|y,x,\theta) = p_{ij} = \frac{\pi_j f(y|x,\beta_j,\sigma_j^2)}{\sum_{j=1}^{J} \pi_j f(y|x,\beta_j,\sigma_j^2)}. \tag{1}$$

This expression combines the regime probabilities π_j with the individual likelihood (f, based on the current estimates for the β_j and σ_j^2) for each regime. The estimates $\hat{p}_{ij}$ yield updated values for $\hat{\pi}_j$ as the average of the posterior probabilities:

$$\hat{\pi}_j = (1/n) \sum_i \hat{p}_{ij}.$$

With the posterior probabilities in hand, each observation can be assigned to the regime with the highest associated probability, in a so-called hard assignment.

The estimation of the model parameters is complex and involves a log-likelihood function that contains both the regime probabilities as well as the regime regression parameters. For independent

[4]See also Leisch (2004) and Grün and Leisch (2008) for a detailed exposition.

observations, this boils down to the familiar sum of the individual log-likelihoods:

$$\log L = \sum_{i=1}^{n} \log(\sum_{j=1}^{J} \pi_j f(y|x, \beta_j, \sigma_j^2).$$

However, the joint estimation of all parameters is not possible. In practice, the EM algorithm of Dempster et al. (1977) is used. This consists of an iteration between the estimation of regime membership probabilities $\hat{\pi}_j$, the so-called *expectation* step, and the estimation of the regression coefficients, the *maximization* step. The latter is based on a weighted log-likelihood, using the estimates of the regime probabilities to weight each observation.

Given estimates for each $\hat{\pi}_j$ and the regime regression parameters β_j and σ_j^2, Equation (1) can be used to compute posterior probabilities that each observation i belongs to a regime j, or $\hat{p}_{ij}$. These probabilities are then used in a weighted maximization process to obtain estimates for each β_j and σ_j^2:

$$\max_{\beta_j, \sigma_j^2} \sum_{i=1}^{n} \hat{p}_{ij} \log f(y_i|x_i, \beta_j, \sigma_j^2).$$

In addition, they also yield updated values for the $\hat{\pi}_j$. Iteration between the two steps continues until convergence is reached.

The resulting hard assignment does not typically yield proper spatial regimes, since it is difficult to integrate a spatial constraint in the mixture estimation. In the literature, there have been a few attempts to introduce spatial aspects into this framework, such as Wall and Liu (2009), Lee (2018), and Bolin et al. (2019). However, these tend to either be based on a geostatistical and Bayesian perspective, or require a panel data structure, neither or which are within the scope of the current paper.

In our application, we use the R package `flexmix` to implement the finite mixture regressions (Leisch, 2004; Grün and Leisch, 2008).

3.3 Spatially Constrained Endogenous Regimes

The endogenous spatial regime estimation method we introduce in this paper is based on an extension of the SKATER spatially constrained clustering algorithm. This approach utilizes a graph partitioning logic in a divisive hierarchical clustering application. The key concept is the creation of a graph structure that enforces contiguity, in the sense that only nodes (observations) that are spatial neighbors have an edge between them in the graph.

The point of departure is a $n \times n$ spatial weights matrix, $\mathbf{W}$, created using one of the common conventions (e.g., contiguity, distance bands, k-nearest neighbors).[5] The weights matrix provides the structure for a graph, where each observation is a node and the edges correspond to non-zero elements in the spatial weights matrix.

The edge weight is based on the attribute similarity between the pair of observations, computed as a squared Euclidean distance. For example, with a $n \times k$ (standardized) matrix $\mathbf{Z}$ of observations on the relevant variables, the distance squared between two contiguous spatial units i and u is:

$$d(i,u) = d(z_i, z_u) = \sum_{p=1}^{k} (z_{ip} - z_{up})^2,$$

where z_i and z_u are $k \times 1$ vectors of observations.

The resulting weighted graph is then reduced to a minimum spanning tree (MST), i.e., such that there is a path that connects all observations (nodes), without any cycles (circular paths) and such that the sum of the edge weights is minimized. In other words, the n nodes are connected by $n-1$ edges, such that the overall between-node dissimilarity is minimized.

At this point, our approach deviates from the standard SKATER implementation. In SKATER, the MST is *pruned* by dropping an edge, i.e., cutting the connection between two observations, such that the between group dissimilarity is decreased the most. To accomplish this, each potential split is evaluated in terms of its contribution to the objective function. Since the graph structure is based on contiguity, each resulting subgraph also consists of spatially contiguous entities.

In our endogenous regimes implementation, the objective function is no longer the between

[5] See Anselin and Rey (2014), Chapters 3 and 4, for an extensive discussion.

group similarity, but instead is based on the fit of the regression, i.e., the sum of squared residuals (SSR). The SSR is readily computed from the difference between the observed and predicted values:

$$\text{SSR} = \sum_i (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})$$

where the predicted value $\hat{\mathbf{y}}$ is obtained for each particular model.

More precisely, in a standard OLS regression, the predicted value is the usual $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$. In the spatial lag model, the predicted value is obtained from the reduced form, as:

$$\hat{\mathbf{y}} = (\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}\mathbf{X}\hat{\beta}.$$

In the spatial error model, the spatial autoregressive parameter is not involved in the predicted value itself, only in its precision. We can gain a computational advantage by exploiting the unbiasedness property of OLS for this specification, so that we can base our algorithm on the SSR of a simpler regression, using OLS. After the regimes have been obtained, the proper FGLS method can be applied to obtain the spatial parameters.

Our heuristic proceeds in the same fashion as SKATER, but now using the SSR for each tree as the criterion to select the optimal pruning of the MST. Formally, for each tree T, the removal of an edge E results in two subtrees, say a and b. The decrease in SSR that results from that cut is:

$$f(E) = \text{SSR}_T - (\text{SSR}_a + \text{SSR}_b),$$

where SSR_a and SSR_b are the contributions of each subtree. The edge is pruned for that E where the difference $f(E)$ is the largest. At this point, we repeat the process for the new set of subtrees to select an optimal cut. This is continued until the desired number of regimes (J) is reached.

In sum, if the objective is to obtain J spatial regimes, $J - 1$ edges must be removed. The iterative process can be described as:

1. Run the econometric regression considering all spatial units to compute the total sum of squared residuals (SSR_T);

2. For each edge E in the MST, compute the objective function $f(E)$ to identify the edge with the highest $f(E)$;
3. Remove E, creating two sub-trees;
4. As long as the number of sub-trees is smaller than the desired number of clusters, repeat 2 and 3 evaluating the objective function based on the total SSR of each sub-tree and the edges it contains.

At the end of the process, the regime regression can be re-run to obtain standard errors and other model diagnostics.

One drawback of this approach (and of SKATER in general) is that it is computationally demanding, since the objective function needs to be evaluated for every potential cut. In addition, since it is a hierarchical method, once an observation is assigned to a particular subtree, it cannot be swapped to a different subtree. Also, in order to be suitable in a regression context, a minimum size constraint must be imposed on each subtree, to assure reliable coefficient estimates. Finally, as with any heuristic, there is no guarantee that an optimal solution is obtained.

The method, which here we label Skater-regression, or Skater-reg for short, is implemented as part of the `spreg` module in PySAL, a Python-based library for spatial data analysis (Rey and Anselin, 2007; Rey et al., 2021).

4 Empirical Illustration

4.1 Data

The data we used to illustrate various regime regression approaches come from a well-known Kaggle set that contains information on house sales between May 2014 and May 2015 in King County, Washington, which includes the city of Seattle.[6]

The original data set contains 21,613 observations as points with coordinates as latitude and longitude, which we reprojected to UTM Zone 10N to yield distance measures as meters. From this

[6]The dataset is available from `https://geodacenter.github.io/data-and-lab//KingCounty-HouseSales2015/`

original set, we removed 353 repeat sales and only kept information on the latest transaction at that location. We further removed 1,147 observations that had multiple sales at the same location which were not repeat sales, suggesting a multi-unit structure. In addition, we dropped 254 transactions that were more than 500m from their nearest neighbor, as well as 14 remote locations in the eastern part of the county.

We further cleaned the data by removing obvious coding errors, such as an observation with 33 bedrooms, sales with zero bedrooms or zero bathrooms, as well as sales on Vashon and Maury islands (these created problems with the spatial weights used in the clustering routines). The final data set contains 19,687 observations.

4.2 Hedonic Model Specification

The regression specification uses log10 of the price (`logprice`) as the dependent variable to control for the extreme skewness of the price variable, as is customary in hedonic model applications. We include all continuous house characteristics available in the Kaggle data set as explanatory variables. In addition, we also computed `age` (and age squared – `age2` –, to allow for the typical nonlinear association between price and age) from the year built variable and created new indicator variables as groupings of the original categories, to control for the lack of observations in individual categories. We also included a completely new variable to proxy density and sprawl in the form of the distance to the nearest neighbor (`distn`). This was calculated from the house coordinates. The full list of variables and their definition is given in Table 1, with descriptive statistics in Table 2. A general impression of the spatial distribution of the log10 of sales prices is presented in the box map in Figure 1. This reveals an overall trend of higher prices closer to the core of Seattle and near the waterfront. Of the 19,687 observations, only 270 turn out to be upper outliers and only 38 lower outliers.

As a reference, the results of an ordinary least squares regression of the basic specification are listed in Table 3. Except for the number of bedrooms, which has a negative sign, all the signs are in the expected direction. The effect of distance to the nearest neighbor is negative, suggesting a minor premium for higher density. All coefficients are highly significant, except for `sqft_lot15`.

Also `renovated` and `sqft_lot`, while still strongly significant, are less so than the other coefficients. The overall fit amounts to an R^2 of 0.655.

4.3 Regime Delineation

We consider six different regime specifications, three of which are spatially contiguous – zip code zones, classic SKATER clusters, and our spatially constrained endogenous regimes –, and three that are not necessarily spatially contiguous – K-means clustering, K-means with x-y coordinates, and the finite mixture model.[7]

As an example of an exogenous regime, we construct a spatial aggregation of the zip code zones that cover King county into five regions, utilizing K-means clustering on the centroid coordinates of the zip code zones. We chose five regions, since this turned out to be the best selection of number of clusters in the spatially constrained endogenous regimes. To illustrate the different methods, we used the same value for all regime specifications to allow for easy comparison, even though we recognize that this number of clusters may not be optimal in each particular case.

The resulting number of observations by category for the zip regimes are listed in the fourth column of Table 4 and the corresponding spatial layout is shown in panel (a) of Figure 2.

Note that the regime label as such has no meaning. We follow the convention used in GeoDa to label categories in decreasing order of the number of observations. For the zip code regions, the regimes range from 7383 to 977 observations, with the last category about a third the size of the previous two. The first two categories are substantially larger. The spatial layout is characterized by a clear west-east split, with the former consisting of three north-south categories, and the latter of two.

Three data-driven regimes are based on the results of, respectively, K-means clustering, K-means clustering with x,y coordinates included, and SKATER spatially constrained clustering. The clustering is carried out on the nine continuous explanatory variables (not including `floors`). For the application of SKATER, a symmetric k-nearest neighbor spatial weights matrix is used with k=17. This ensured that all observations were spatially connected, which is required for the

[7] All cluster results are obtained with the GeoDa software (Anselin et al., 2021).

proper application of the algorithm.

The quality of the clusters computed with the different methods varies greatly. The best result in terms of the ratio of between sum of squares to total sum of squares is obtained for K-means (0.5186). However, the introduction of the x-y coordinates in the feature set lowers this to 0.4527. Finally, the imposition of the spatial constraint in SKATER pays a heavy price in terms of intra-cluster similarity, with the ratio decreasing to 0.1168. In other words, in order to obtain spatially compact regimes, the within-cluster similarity on the nine continuous variables declines considerably.

The corresponding number of observations by regime are listed in columns 1, 2 and 5 of Table 4. The distribution among the five categories is quite different for each of the cluster results. SKATER yields the largest single group, with 7766 observations, but both K-means results have a very small fifth category, with respectively 283 and 288 observations (SKATER's smallest category has 1842). The top four categories for K-means x-y are fairly evenly balanced (ranging from 5188 to 4359), whereas this is the case for SKATER's three middle categories (3618 to 3080). K-means is dominated by the two largest groups (6870 and 6092 observations). The spatial layout of the SKATER result is shown in panel (b) of Figure 2. We see three north-south stacked regimes in the western part of the county, with two large north-south regions to the east. The largest region is in the center of the map.

The spatial layout for the non-spatial clusters is not shown, since it has no particular spatial organization and any pattern would be hard to distinguish on a point map with so many observations. Instead, we proxy the extent to which the compactness of a regime approaches that of the spatially contiguous regimes by computing the ratio of join counts of the same regime classification to the total number of joins in each regime. For the spatially contiguous regimes, this ratio should be equal to one, except for boundary effects. For the others, the higher the ratio, the more the spatial layout approaches one where an observation in a given category is surrounded by neighbors in the same category.

The results are summarized in Table 5. For K-Means and K-Means (x-y), the ratios are listed in the first two columns. As is to be expected, K-Means (x-y) does better than K-Means, since it includes the coordinates in the clustering objective. It achieves an overall ratio of 0.6689, compared

to 0.5396 for K-Means. Only one of the regimes has a ratio below 0.5 and the highest value is 0.8064. All the individual regime ratios for K-Means (x-y) clearly dominate K-Means.

The cluster allocation for the finite mixture model and the spatially constrained endogenous regimes, here labeled as Skater-regression, are obtained as part of the estimation process. For comparison purposes, the number of observations by regime for a hard allocation in the finite mixture model and the outcome for the Skater-regression are given in columns 3 and 6 of Table 4. The finite mixture results show a fairly linear decline from 6045 in the largest group to 823 in the smallest. For Skater-Reg, the smallest group has 336 observations, similar to the results for the two K-means approaches, but only about one sixth of the size for the SKATER cluster method. The regimes are dominated by two larger groups (7425 and 6261) and evenly balanced remaining categories. The spatial layout of Skater-reg, shown in panel (c) of Figure 2 is quite distinct of that of the other two spatial methods in that much less north-south stacking is present. The two largest regions span from west to east, one in the north and middle, the other in the southern part of the county. The smallest region is an enclave situated in the city of Bellevue, on the east coast of Lake Washington. In contrast, the join count ratio for the finite mixture results (in the third column of Table 5) are the worst, with an overall ratio of 0.4270, well below 0.5, and the individual regime ratios ranging from 0.2570 to 0.4851. Clearly, in this particular empirical application, the optimization of the regime regression fit is obtained by ignoring any spatial structure.

4.4 Regime Regression Results

The six regime classifications, with five categories each, yield a total of 30 sets of 17 regression coefficients. It is beyond our scope to delve into these in detail, since our focus is on the role of the regime selection. The results are summarized in Table 6, which lists for each variable and each regime classification those categories where the coefficient was *not* significant at $p < 0.01$. Recall that the category labels have no meaning and only indicate the relative size of each category (with 1 having the largest number of observations). We chose this summary since all variables except `floors` are significant in the majority of cases (i.e., more than 15). In addition to the constant term, the variable `sqft_liv` is consistently significant in all cases. A few others, such as `sqft_liv15`, the

`grade` indicator variables, and the `view` indicator variable only fail to be significant in a handful of situations (less than five out of thirty).

Other variables with more than 20 instances of significant coefficients include `age2`, `condn`, `sqft_lot`, `bath`, and `age`. The evidence for the remaining variables is mixed, alternating significance in some regimes with lack thereof in others.

In all, each regime classification yields 85 estimated coefficients. Across this broad spectrum, the finite mixture model achieves the best result, with 79 significant results (6 non significant in Table 6), which is expected, since it optimizes the fit of the submodels. The two K-means approaches give the worst result, with respectively 58 (K-means x-y) and 59 (K-means) significant coefficients. The Skater-regression endogenous regimes yield 64 significant coefficient estimates. Whereas this is only the fourth best result, the Skater-reg is second only to the finite mixture model in terms of overall fit, with an R^2 of 0.8379 compared to 0.9226 for finite mixtures. This is much better than the data-driven K-means (0.6745 and 0.7609) and SKATER results (0.7545). Intriguingly, the ad hoc zip code regions are third both in terms significant coefficients as well as overall fit, with 67 significant coefficients and an R^2 of 0.8086.

The full results for the Skater-regression regimes are listed in Table 7, which contains details on the significance of each coefficient in the respective regimes. The fit in the individual regimes (computed using the regime-specific mean of the dependent variable rather than the overall mean) ranges from 0.857 in the smallest subset (with 336 observations) to 0.693 in the second largest (n = 6261). The fit for the other three regimes is around 0.75.

Finally, we show the outcome of the Chow test on constancy of the regression coefficients across regimes in Table 8. The overall test rejects the null hypothesis on constancy with very high significance in all regimes settings. However, only for the finite mixture results is the rejection also for each individual coefficient at $p < 0.001$. Using $p < 0.05$, Skater-reg also achieves this result. For the other regimes SKATER fails to reject (at $p < 0.05$) for `sqft_lot` and `bedrooms`, Zip for `bathrooms`, K-means for `sqft_liv` and K-means (x-y) for `sqft_liv` and `floors`. The dominant result, however, is rejection of the null hypothesis, suggesting (strong) regime heterogeneity.

Overall, the Skater-regression comes closest in terms of fit to the gold standard of the finite

mixture regression. Whereas the latter yields a set of collections of individual observations that have no spatial structure whatsoever, our new approach results in spatially cohesive sub-regions. The loss in overall fit relative to the finite mixture result is the price to pay to achieve the objective of spatial contiguity.

5 Concluding Remarks

In this paper, we proposed a new approach to determine spatial regimes endogenously by incorporating the regression fit in the objective function of a spatially constrained regionalization method. The results are encouraging, yielding both a good overall fit and meaningful spatial subregions for the observations in our empirical illustration. Clearly, further experimentation is needed to address the performance of our approach in other settings.

In our empirical application, the data driven methods to obtain regime definitions did not do well in terms of fit. This may be due to the peculiarities of the particular data set, but it is something to keep in mind. It suggests that it is important to assess the sensitivity of the results and the interpretation of the associated regimes for more than one method.

The finite mixture approach does best, confirming earlier findings in the literature. However, the resulting regimes are simply collections of individual observations that yield the best fit and typically do not have a meaningful spatial interpretation. When the latter is an objective, as in the case of *spatial* regimes, our preliminary results suggest that the Skater-regression approach provides a viable alternative. The difference in fit provides an indication of the extent of the trade off required to satisfy the spatial constraint.

References

Anselin, L. (1988). *Spatial Econometrics: Methods and Models*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. (1990). Spatial dependence and spatial structural instability in applied regression analysis. *Journal of Regional Science*, 30:185–207.

Anselin, L. and Arribas-Bel, D. (2013). Spatial fixed effects and spatial dependence in a single cross-section. *Papers in Regional Science*, 92:3–17.

Anselin, L., Li, X., and Koschinsky, J. (2021). GeoDa, from the desktop to an ecosystem for exploring spatial data. Working paper, Center for Spatial Data Science, University of Chicago, Chicago, IL. DOI:10.13140/RG.2.2.33158.09287.

Anselin, L. and Lozano-Gracia, N. (2009). Spatial hedonic models. In Mills, T. and Patterson, K., editors, *Palgrave Handbook of Econometrics: Volume 2, Applied Econometrics*, pages 1213–1250. Palgrave Macmillan, Basingstoke, United Kingdom.

Anselin, L. and Rey, S. J. (2014). *Modern Spatial Econometrics in Practice, A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press, Chicago, IL.

Assunção, R., Lage, J., and Reis, E. (2002). Análise de conglomerados espaciais via árvore geradora mínima. *Revista Brasileira de Estatística*, 63(220):7–24.

Assunção, R. M., Neves, M., Camara, G., and Da Costa Freitas, C. (2006). Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, 20:797–811.

Belasco, E., Farmer, M. C., and Lipscomb, C. A. (2012). Using a finite mixture model of heterogeneous households to delineate housing submarkets. *The Journal of Real Estate Research*, 34:577–594.

Bertsimas, D. and Shioda, R. (2007). Classification and regression via integer optimization. *Operations Research*, 55:252–271.

Bhattacharjee, A., Castro, E., Maiti, T., and ao Marques, J. (2016). Endogenous spatial regression and delineation of submarkets: A new framework with application to housing markets. *Journal of Applied Econometrics*, 31:32–57.

Bolin, D., Wallin, J., and Lindgren, F. (2019). Latent Gaussian random field mixture models. *Computational Statistics and Data Analysis*, 130:80–93.

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

Bourassa, S. C., Cantoni, E., and Hoesli, M. (2010). Predicting house prices with spatial dependence: A comparison of alternative methods. *Journal of Real Estate Research*, 32:139–160.

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

Bourassa, S. C., Hoesli, M., and Peng, V. S. (2003). Do housing submarkets really matter? *Journal of Housing Economics*, 12:12–28.

Brasington, D. M. and Hite, D. (2005). Demand for environmental quality: a spatial hedonic analysis. *Regional Science and Urban Economics*, 35:57–82.

Casetti, E. (1972). Generating models by the expansion method: Applications to geographical research. *Geographical Analysis*, 4:81–91.

Casetti, E. (1997). The expansion method, mathematical modeling, and spatial econometrics. *International Regional Science Review*, 20:9–33.

Chow, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28:591–605.

Cleveland, W. S. and Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Duque, J., Anselin, L., and Rey, S. J. (2012). The max-p-regions problem. *Journal of Regional Science*, 52:397–419.

Elhorst, J. P. (2014). *Spatial Econometrics, From Cross-Sectional Data to Spatial Panels*. Springer, Heidelberg.

Fotheringham, A. S. (1997). Trends in quantitative methods I: Stressing the local. *Progress in Human Geography*, 21:88–96.

Fotheringham, A. S., Brundson, C., and Charlton, M. (1998). Geographically weighted regression: A natural evolution of the expansion method for spatial data analysis. *Environment and Planning A*, 30:1905–1927.

Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2002). *Geographically Weighted Regression*. John Wiley, Chichester.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York, NY.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98:387–396.

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

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

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

Grün, B. and Leisch, F. (2008). FlexMix Version 2.0: Finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28.

Guo, D. (2008). Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). *International Journal of Geographical Information Science*, 22:801–823.

Guo, D. and Wang, H. (2011). Automatic region building for spatial analysis. *Transactions in GIS*, 15:29–45.

Helbich, M., Brunauer, W., Hagenauer, J., and Leitner, M. (2013). Data-driven regionalization of housing markets. *Annals of the Association of American Geographers*, 108:871–889.

Kuminoff, N. V., Parmeter, C. F., and Pope, J. C. (2010). Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities? *Journal of Environmental Economics and Management*, 60:145–160.

Lee, J. (2018). A spatial latent class model. *Economics Letters*, 162:62–68.

Lee, L.-F. and Yu, J. (2011). Estimation of spatial panels. *Foundations and Trends in Econometrics*, 4:1–164.

Leisch, F. (2004). FlexMix. a general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11.

Lloyd, C. D. (2010). *Local Models for Spatial Analysis, Second Edition*. CRC Press, Boca Raton, FL.

Loader, C. (1999). *Local Regression and Likelihood*. Springer-Verlag, Heidelberg.

McLachlan, G. and Basford, K. (1988). *Mixture Models: Inference and Applications to Clustering*. Marcel Dekker, New York, NY.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York, NY.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual Review of Statistics and Its Application*, 6:355–378.

Olson, A. W., Zhang, K., Calderon-Figueroa, F., Yakubov, R., Sanner, S., Silver, D., and Arribas-Bel, D. (2021). Classification and regression via integer optimization for neighborhood change. *Geographical Analysis*, 53:192–212.

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

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

Raudenbush, S. W. and Bryk, A. S. (2002). *Hierarchical Linear Models (Second Edition)*. Sage Publications, Newbury Park, CA.

Rey, S., Anselin, L., Amaral, P., Arribas-Bel, D., Cortes, R., Gaboardi, J., Kang, W., Knaap, E., Li, Z., Lumnitz, S., Oshan, T., Shao, H., and Wolf, L. (2021). The PySAL ecosystem: Philosophy and implementation. *Geographical Analysis*. DOI:10.1111/gean.12276.

Rey, S. J. and Anselin, L. (2007). PySAL: A Python library of spatial analytical methods. *The Review of Regional Studies*, 37:5–27.

Ugarte, M., Goicoa, T., and Militino, A. (2004). Searching for housing submarkets using mixtures of linear models. In LeSage, J. P. and Pace, R. K., editors, *Advances in Econometrics: Spatial and Spatiotemporal Econometrics*, pages 259–279. Elsevier Science Ltd., Oxford, UK.

Wall, M. M. and Liu, X. (2009). Spatial latent class analysis model for spatial distributed multivariate binary data. *Computational Statistics and Data Analysis*, 53:3057–3069.

Wolf, L. J., Oshan, T. M., and Fotheringham, A. S. (2018). Single and multiscale models of process spatial heterogeneity. *Geographical Analysis*, 50:223–246.

Table 1: Variable Definitions

Variable	Definition
price	sale price (dependent variable used as log10)
bedrooms	number of bedrooms
bathrooms	number of bathrooms
sqft_liv	size of living area in square feet
sqft_lot	size of the lot in square feet
floors	number of floors
renovated	1 if renovated
age	age of structure (computed from yr_built variable)
age2	square of age
sqft_liv15	average size of closest 15 houses, in square feet
sqft_lot15	average size of the closest 15 houses' lots, in square feet
viewd	1 if view is > 0 on the original view scale of 0 to 4
condn	1 if condition > 3 (better than poor, fair and average)
avggrade	1 if grade $= 7$ (average grade, plats and sub-divisions)
abvavgrd	1 if grade $= 8$ (just above average, better materials)
greatgrd	1 if grade > 8 (better design, high quality, custom, mansion)
distnn	distance to nearest neighbor in meters

Table 2: Descriptive Statistics

variable	mean	std	min	50%	max
price	546,677	374,938	78,000	454,000	7,700,000
log(price)	5.671	0.230	4.892	5.657	6.887
bedrooms	3.40	0.90	1	3	11
bathrooms	1.76	0.73	1	2	8
sqft_liv	2104	918.8	390	1940	12050
sqft_lot	13624	34109	520	7679	1164794
floors	1.4	0.54	1	1	3
renovated	0.043	0.202	0	0	1
age	44.3	29.1	0	41	115
age2	2811	3113.9	0	1681	13225
sqft_liv15	2003.7	689.3	399	1860	6210
sqft_lot15	11414	20027.4	659	7660	411962
viewd	0.099	0.299	0	0	1
condn	0.348	0.476	0	0	1
avggrade	0.417	0.493	0	0	1
abvavgrd	0.280	0.449	0	0	1
greatgrd	0.202	0.401	0	0	1
distnn	97.6	71.2	11.1	78.7	495.9

Table 3: OLS Estimation Results

variable	coefficient	std. error	p-value
constant	5.1179	0.0083	0.0000
bedrooms	-0.0177	0.0014	0.0000
bathrooms	0.0281	0.0021	0.0000
sqft_liv	0.0000897	0.0000022	0.0000
sqft_lot	0.0000001	0.000000	0.0008
floors	0.0183	0.0027	0.0000
renovated	0.0136	0.0051	0.0077
age	0.00091	0.00016	0.0000
age2	0.000015	0.000001	0.0000
sqft_liv15	0.000048	0.000002	0.0000
sqft_lot15	-0.0000001	0.0000001	0.1399
viewd	0.0683	0.0035	0.0000
condn	0.0220	0.0023	0.0000
avggrade	0.1270	0.0036	0.0000
abvavgrd	0.2121	0.0042	0.0000
greatgrd	0.3108	0.0055	0.0000
distnn	-0.00020	0.00002	0.0000
R^2	0.655		

Table 4: Number of Observations By Regime

	Non-Spatial			Spatial		
	K-Means	K-Means (x-y)	Finite Mixture	Zip Zones	Skater	Skater-Regression
1	6970	5188	6045	7383	7766	7425
2	6092	5056	5904	5498	3618	6261
3	3661	4796	4588	2991	3381	2874
4	2681	4359	2327	2838	3080	2791
5	283	288	823	977	1842	336

Table 5: Join Count Ratio for Non-Spatial Regimes

	K-Means	K-Means (x-y)	Finite Mixture
1	0.5612	0.6674	0.4597
2	0.5351	0.6218	0.4851
3	0.5339	0.8064	0.3640
4	0.5164	0.5917	0.3792
5	0.3958	0.4028	0.2570
Overall	0.5396	0.6689	0.4270

Table 6: Regime Estimation Results - Coefficients *Not* Significant at p< 0.01

variable	K-Means	K-Means (x-y)	Fin. Mixture	Zip	Skater	Skater-Reg
constant	–	–	–	–	–	–
bedrooms	4,5	1,3,5	–	3,4,5	3	2,3,5
bathrooms	4,5	1,3,5	5	5	3	–
sqft_liv	–	–	–	–	–	–
sqft_lot	1,2,3,4	2,4	3	1,2	1,2,3	3,4,5
floors	3,4,5	1,2,3,4,5	–	1,2,3,5	4,5	2,5
renovated	2,4,5	1,5	–	–	3,4	3,5
age	3,4	4	–	2,4,5	2,4	1,3
age2	4	–	–	–	2,4	1,4
sqft_liv15	–	–	4	–	–	–
sqft_lot15	1,3,5	2,4,5	3	5	–	–
viewd	5	5	–	5	–	–
condn	5	5	1,5	5	–	5
avggrade	3	4	–	–	–	5
abvavgrd	3	4	–	–	–	5
greatgrd	3	4	–	–	–	–
distnn	5	2,4,5	–	3,4	3,4,5	1,2,3,5
R^2	0.6745	0.7609	0.9226	0.8086	0.7545	0.8379

Table 7: Spatially Constrained Endogenous Regimes – OLS Estimation

	Regime 1	Regime 2	Regime 3	Regime 4	Regime 5
CONSTANT	5.3286***	5.1639***	5.345***	5.1768***	5.7688***
	(0.01)	(0.0107)	(0.0142)	(0.0168)	(0.0506)
bedrooms	-0.0047***	-0.0025	-0.0006	-0.0118***	-0.009
	(0.0016)	(0.0018)	(0.0022)	(0.0028)	(0.0062)
bathrooms	0.0163***	0.006**	0.00938***	0.00987**	0.02391***
	(0.0023)	(0.0027)	(0.0032)	(0.0043)	(0.0083)
sqft_liv	0.0001***	0.0001***	0.0001***	0.0001***	0.0001***
	(0.000002)	(0.000003)	(0.000004)	(0.000005)	(0.000008)
sqft_lot	0.0000001***	0.0000004***	0.000002	0.000001	-0.000001
	(0)	(0)	(0.000001)	(0.000001)	(0.000001)
floors	-0.0089***	0.0037	-0.0106***	0.0175***	-0.0115
	(0.0029)	(0.0035)	(0.0041)	(0.0062)	(0.013)
renovated	0.067***	0.0636***	0.0031	0.0199**	0.0309
	(0.0069)	(0.007)	(0.0072)	(0.0089)	(0.0179)
age	0.0001	-0.0014***	-0.0001	0.0017***	-0.0035***
	(0.0002)	(0.0002)	(0.0003)	(0.0003)	(0.0009)
age2	-0.000001	0.000012***	0.00001***	-0.000004	0.000033***
	(0.000002)	(0.000002)	(0.000002)	(0.000003)	(0.000008)
sqft_liv15	0.00004***	0.00004***	0.00007***	0.00009***	0.00003***
	(0.000003)	(0.000003)	(0.000005)	(0.000006)	(0.000009)
sqft_lot15	-0.0000003***	0.0000003***	-0.000005***	-0.00001***	0.00001***
	(0.0000001)	(0.0000001)	(0.000001)	(0.000001)	(0.000002)
viewd	0.0841***	0.098***	0.0264***	0.0392***	0.0631***
	(0.0039)	(0.0048)	(0.0058)	(0.0068)	(0.014)
condn	0.0489***	0.0246***	0.0122***	0.0487***	0.0079
	(0.0026)	(0.0026)	(0.0039)	(0.0052)	(0.0112)
avggrade	0.0579***	0.0417***	0.0718***	0.0899***	-0.0236
	(0.0057)	(0.004)	(0.0064)	(0.0066)	(0.0333)
abvavgrd	0.1166***	0.0746***	0.1173***	0.1871***	0.003
	(0.0062)	(0.0051)	(0.0076)	(0.0087)	(0.034)
greatgrd	0.1782***	0.1109***	0.2139***	0.3118***	0.073**
	(0.0071)	(0.0068)	(0.01)	(0.0117)	(0.0368)
distnn	0.00002	0.00001	-0.00002	-0.00012***	0.00008
	(0.000016)	(0.000016)	(0.000039)	(0.000047)	(0.000086)
n	7425	6261	2874	2791	336
R^2	0.765	0.693	0.752	0.749	0.857

Note 1: Standard deviation in parenthesis
Note 2: ** indicate significant at 5% and *** indicate significant at 1%.

Table 8: Chow Test Results - p values

variable	K-Means	K-Means (x-y)	Fin. Mixture	Zip	Skater	Skater-Reg
constant	0.000	0.000	0.000	0.000	0.000	0.000
bedrooms	0.000	0.000	0.000	0.004	0.071	0.021
bathrooms	0.005	0.000	0.000	0.082	0.004	0.022
sqft_liv	0.089	0.237	0.000	0.000	0.000	0.000
sqft_lot	0.007	0.000	0.000	0.000	0.200	0.000
floors	0.007	0.267	0.000	0.004	0.000	0.000
renovated	0.000	0.000	0.000	0.001	0.009	0.000
age	0.000	0.000	0.000	0.000	0.000	0.000
age2	0.000	0.000	0.000	0.000	0.000	0.000
sqft_liv15	0.000	0.000	0.000	0.000	0.000	0.000
sqft_lot15	0.000	0.000	0.000	0.000	0.000	0.000
viewd	0.000	0.000	0.000	0.000	0.000	0.000
condn	0.040	0.000	0.000	0.000	0.000	0.000
avggrade	0.000	0.000	0.000	0.000	0.000	0.000
abvavgrd	0.000	0.000	0.000	0.000	0.000	0.000
greatgrd	0.011	0.000	0.000	0.000	0.000	0.000
distnn	0.000	0.000	0.000	0.000	0.000	0.042
Overall	0.000	0.000	0.000	0.000	0.000	0.000

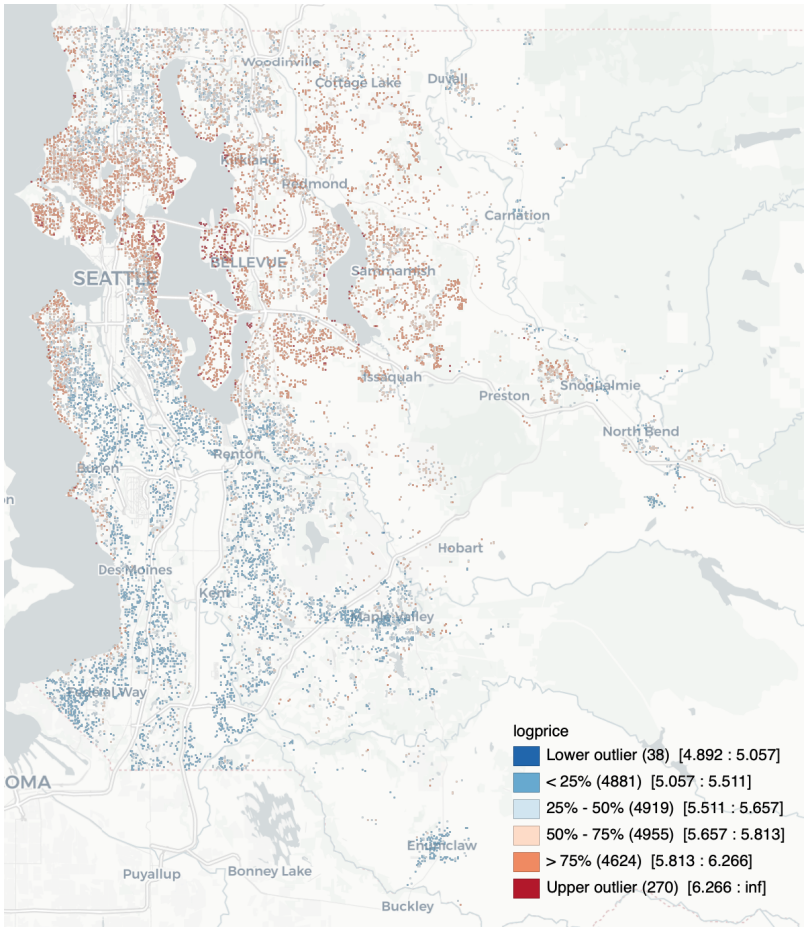

Figure 1: Log of house sale prices in King County between May 2014 and May 2015

Figure 2: Spatial Regimes

(a) Zip

(b) Skater

(c) Skater-Reg