SCHOOL OF ECONOMICS AND MANAGEMENT
FACULTY OF SOCIAL SCIENCES
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

CREATES Research Paper 2010-53

Detecting Housing Submarkets using Unsupervised Learning of Finite Mixture Models

Christos Ntantamis

School of Economics and Management
Aarhus University
Bartholins Allé 10, Building 1322, DK-8000 Aarhus C
Denmark

Detecting Housing Submarkets using Unsupervised Learning of Finite Mixture Models

Christos G. Ntantamis *
Aarhus University, CREATES

August 18, 2010

Abstract

The problem of modeling housing prices has attracted considerable attention due to its importance in terms of households' wealth and in terms of public revenues through taxation. One of the main concerns raised in both the theoretical and the empirical literature is the existence of spatial association between prices that can be attributed, among others, to unobserved neighborhood effects. In this paper, a model of spatial association for housing markets is introduced. Spatial association is treated in the context of spatial heterogeneity, which is explicitly modeled in both a global and a local framework. The global form of heterogeneity is incorporated in a Hedonic Price Index model that encompasses a nonlinear function of the geographical coordinates of each dwelling. The local form of heterogeneity is subsequently modeled as a Finite Mixture Model for the residuals of the Hedonic Index. The identified mixtures are considered as the different spatial housing submarkets. The main advantage of the approach is that submarkets are recovered by the housing prices data compared to submarkets imposed by administrative or geographical criteria. The Finite Mixture Model is estimated using the Figueiredo and Jain (2002) approach due to its ability in endogenously identifying the number of the submarkets and its efficiency in computational terms that permits the consideration of large datasets. The different submarkets are subsequently identified using the Maximum Posterior Mode algorithm. The overall ability of the model to identify spatial heterogeneity is validated through a set of simulations. The model was applied to Los Angeles county housing prices data for the year 2002. The results suggests that the statistically identified number of submarkets, after taking into account the dwellings' structural characteristics, are considerably fewer that the ones imposed either by geographical or administrative boundaries.

JEL Classification Numbers: C13, C21, R0

Keywords: Hedonic Models, Finite Mixture Model, Spatial Heterogeneity, Housing Submarkets

[*] Address: School of Economics & Management, University of Aarhus, Building 1322, 8000 Aarhus C, Denmark, telephone: +45 89422140, email: *cntantamis@creates.au.dk*. This paper constitutes the first part of my PhD thesis in McGill University. I would like to express gratitude to Professor Leah Brooks for providing me with the data and for her suggestions, and to Professor John W. Galbraith for helpful comments. I also benefited from invaluable comments from the participants of the Society for Computational Economics International Conference in Montreal, of the Canadian Economics Association Annual Meeting,in Halifax, and the Department of Economics Ph.D. Seminars, McGill University. Any remaining errors are my own. Center for Research in Econometric Analysis of Time Series, CREATES, is funded by The Danish National Research Foundation.

1 Introduction

The problem of modeling dwelling selling prices is still a challenge for housing analysts, especially in view of the importance that real estate wealth has in the household's consumption and thus in the entire economy. Hedonic market models, which are observed on the equilibrium between demand and supply at a given time, have been used for this purpose as a common modeling approach. Hedonic models (or indexes) consider the features of an asset to be the determinants of its actual values. Consequently, a house price can be estimated by regressing on the number of rooms, the total living area, the number of bathrooms, and location characteristics. However, it is very difficult to get a complete account of all the relevant location characteristics, so the regression residuals are often found to be spatially correlated.

The latter issue emerges as a direct consequence of the nature of housing compared to other commodities. In particular, housing differs in terms of its high cost, its durability, its heterogeneity, and most importantly its locational fixity. The intrinsic uniqueness of each location gives rise to the spatial heterogeneity property of housing.

Geographic data are predominantly dependent. As Tobler (1970) suggests in his First Law of Geography

> "Everything is related to everything else, but near things are more related than distant things"

This spatial heterogeneity needs to be modeled in order not only to improve the efficiency of the hedonic models estimates, but also to incorporate this spatial information in order to construct more effective models for housing valuations. In the literature there are two approaches to modeling spatial heterogeneity. The first one attempts to model directly the autocorrelation structure of the hedonic model residuals. Dubin (1992,1998) makes use of kriging in order to model the residuals; their correlation is written as a function of the distance between two dwellings [1]. In a similar fashion, Basu and Thibodeau (1998) model the spatial autocorrelation of the residuals using semi-variograms. On the other hand, Can (1992), Can and Megbolugbe (1997), and Dubin et al. (1999), building on Anselin's (1988) model of spatial autocorrelation, incorporate spatially lagged variables directly in the hedonic regression. The estimation of these models requires the construction of weight matrices, which are square matrices with dimension equal to the number of the observations so as to state the spatial autocorrelations, and their subsequent inversion. When the number of observations becomes large, inverting the weight matrix becomes problematic even though this effect is mitigated by the fact that the matrix is sparse.

The second approach to modeling spatial heterogeneity builds on the presumption that we can segment housing markets to clusters of dwellings that need to satisfy the following three conditions[2]: (a) homogeneity, (b) contiguity, and (c) parsimony. Each cluster should constitute a different housing submarket. Thus, the hedonic model, when estimated for dwellings within each cluster, may be able to provide residuals that are spatially uncorrelated. This presumption has its origins in the early housing markets literature. Strazsheim (1974) states:

> "The urban housing market" is a set of compartmentalized and unique submarkets with demand and supply influences likely to result in a different stucture of prices in each."

[1] Kriging corresponds to a set of geostatistical techniques to interpolate the value of a random field at an unobserved location from observations of its value at nearby locations.

[2] Goodman (1981).

Housing markets may be away from the long-run equilibrium as a result of segmented demand or inflexible supply adjustment processes. The standard long-run analysis assumes instantaneous adjustment of housing size, location, quality and distribution of dwelling units to changes in income, employment, location, population, tastes, and transportation (see Strazheim 1975). On the demand side, however, it has been noticed that prospective owners and renters often examine housing in a limited geographical area because of search costs, racial discrimination, or desired proximity to friends or workplace. On the supply side, there are also constraints; capital stock is difficult to modify, and vacant land is scarce especially in urban regions.

The first part of the paper follows the approach of Goodman (1981) and thus it models the spatial heterogeneity in housing prices in the submarkets setting. This approach emphasizes classification accounting for spatial heterogeneity, by making groups of observations define potential submarkets. In contrast to a large part of the literature that defines the location of the submarkets *a priori*, based on socioeconomic considerations, political jurisdictions, school districts or market areas as perceived by the real estate agencies (for example Goodman 1981, Adair et al. 1996, and Goodman and Thibodeau 1998), I estimate the number and the location of the submarkets endogenously. That is, instead of determining the submarkets based on some prior view, an alternative approach that allows the data to speak for themselves is followed. Examples from the literature following the same approach include, among others, Clapp et al. (2002), Ugarte et al. (2004) and Clapp and Wang (2005).

This method allows the researchers to model rather than impose submarkets. The number and the corresponding locations of the submarkets will be determined in an optimal way, i.e. in terms of the data's statistical properties, in order to identify the areas with similar demand and supply functions. Thus, it is expected that the model will identify housing submarkets in a more parsimonious way compared with the determination of submarkets solely in terms of geographically determined or administratively imposed regions. What may appear to be diverse submarkets could in fact be artificial divisions.

In this paper, I propose a model that consists two modules. The first module involves an extended version of the hedonic price model; the geographical coordinates of the dwelling are included to the set of the explanatory variables, as in Clapp and Wang (2005). The hedonic model is estimated using a standard OLS procedure. This yields prices that are standardized for the dwelling's structural characteristics. The second module assumes that the residuals from the first module follow a *Finite Mixture Model*, with each of the mixing components corresponding to a housing submarket. I consider this approach as intuitively appealing; there is a one-to-one correspondence between the mixtures and the housing submarkets.

Similar problems are addressed in engineering applications such as image processing or pattern recognition and as a result a rather voluminous literature has been formed over the past years. The methods that are proposed in this literature present desirable properties both in terms of estimation accuracy and computation efficiency.

I use the Minimum Description Length criterion variant proposed by Figuereido and Jain (2002) for the estimation of the Finite Mixture Model. The particular approach is selected for these merits: a) the number of the mixtures, and thus the number of the housing submarkets, are endogenously identified during the estimation, b) the model estimation algorithm is simple in its implementation since it includes a variant of the standard Expectation Maximization algorithm, and c) it is computationally more efficient, thus allowing the examination of considerably larger data sets (100000 dwellings or more), which makes the methodology appropriate for mass assessment applications [3].

The final step of the modeling procedure is the re-estimation of the hedonic model for each of the identified spatial submarkets. The parameters of the index can be subsequently used in order to obtain the value for any dwelling that belongs to that submarket.

Nowhere in the previous literature were Monte Carlo simulations used to validate the ability of the models to identify spatial clusters. Here, this evaluation of the validity of the model is undertaken with the result indicating that the model can be used for the identification of housing submarkets and for the explanation of the price variation, in data sets containing a large number of observations.

Finally, the model is employed in order to scrutinize the housing market of the Los Angeles county using values of houses sold in year 2002. The sample size obtained is considerably larger compared to what is used in past empirical applications (approximately 108000 dwellings). The results suggest that the statistically identified number of submarkets, after taking into account the dwellings' structural characteristics, are considerably fewer by far than the ones imposed either by geographical or administrative boundaries.

The remaining chapter is structured as follows: Section 2 describes the model used in the analysis in detail. The hedonic price model is reviewed and the estimation procedure for the finite mixture model is introduced. Section 3 discusses the simulation results. In section 4, the results from the application of the model to the real estate market of Los Angeles are presented. The chapter ends with some concluding remarks.

2 The model

2.1 The hedonic model

Since the seminal work of Adelman and Griliches (1961), who provided the rational for the construction of hedonic indices as a quantification of heterogeneous products' quality, a vast literature for the determination of housing prices has emerged. Rosen (1974) provided the theoretical framework for the hedonic models as an approach to the determination of housing prices. He argued that in equilibrium between demand and supply, the price of the house can be determined as a function of its characteristics and the weights can be estimated by running a standard regression. Moreover, Rosen (1974) and Epple (1987) discussed the issues involved in how to obtain the demand and/or supply functions from the estimated hedonic model.

The hedonic price model can be written as

$$\mathbf{y} = \mathbf{X} \cdot \beta + \epsilon \tag{1}$$

where $\mathbf{y}$ is the dependent variable related with the dwelling price and $\mathbf{X}$ corresponds to the set of the dwelling characteristics such as lot size, number of bedrooms, existence of fireplace, neighborhood amenities, etc. The model is estimated by Ordinary Least Squares.

Despite its age, the hedonic model approach is still very popular for housing market valuation. It is preferred over the usual repeat-sales approach used by appraisers as it has been demonstrated (Meese and Wallace 1997, Pace et al 2002) that: a) it actually yields smaller errors in terms of forecasting the realized

[3] Data sets in the existing literature involve a number of housing prices in the vicinity of 1000, for example Adair et al. (1996) consider 1080 transactions, Clapp and Wang (2005)'s sample was of 1069 transactions.

market price, and b) it enables a rather costless evaluation for every new property once the parameters of the hedonic model have been estimated.

There are several issues regarding the specification of the hedonic model, as discussed extensively in Freeman (1979), Mark and Goldberg (1988), Meese and Wallace (1997), Basu and Thibodeau (1998), Reichert (2002), and Pace et al. (2002) but the most important are related to the functional form and to variable selection priority.

Rosen (1974) did not preclude any particular form for the function of the characteristics; the hedonic function could very well take a nonlinear form. The literature addresses the problem of the functional form by considering a Box-Cox transformation of the data. The most usual functional formulations are the semi-log (the log of the price is the dependent variable, whereas the structural characteristics remain unaltered), or the log-log (all variables in the regression function are taken in logs). In this paper the semi-log form is employed for several reasons: a) it is used in the majority of the recent literature, b) it decreases the heteroskedasticity in the residuals, and c) it provides a neat interpretation of the estimated coefficients of the regressions (rate of change in the price for an additional unit of the structural variable).

The dwelling characteristic variables can be separated into two categories. The first one corresponds with the structural characteristics of the dwelling such as age, lot size, number of rooms, existence of pool, etc that can be measured with relative accuracy. The second one corresponds with the neighborhood characteristics such as racial composition, amenities, environmental aspects, etc, which generate the most problems. Modeling all the neighborhood effects in a coherent way is challenging due to difficulties in identifying the different effects, and in constructing proper quantitative variables for their description. The problem of omitted variables which will result in biased and inconsistent estimates for the model will inevitably emerge. Moreover, quantifying the quality aspects of the particular neighborhood is very cumbersome as it will require the construction of detailed questionnaires to be completed by a significant number of objective evaluators (Kain and Quigley 1970). Finally, such a task is simply implausible for the case of Mass Assessment when the task involves estimating the value of all dwellings in a large city.

In order to circumvent these issues, neighborhood characteristics will not be included in the analysis as it has been argued before that the explicit modelling of the spatial heterogeneity can capture these effects.

Lastly, a “global” form of heterogeneity is incorporated in the hedonic model by encompassing a non-linear function of the geographical coordinates of each dwelling. The acquisition of the coordinates is made possible by the use of geocoding (a Geographic Information Systems (GIS)application that allows the exact pinpointing of each dwelling, based on its address, in terms of its geographical coordinates). The coordinates function operates as a proxy for location-specific information. This method is preferred to the one that simply puts neighborhood dummy variables since it leaves less space for arbitrary decisions such as the selection of the neighborhoods' boundaries [4]. Moreover, since the focus of this paper is in the endogenously identification of submarkets the employment of information of that type is not desirable.

2.2 Finite mixture Model

The hedonic price model takes into account the structural characteristics of the dwelling. The spatial attributes will be estimated by assuming a finite mixture model for the residuals from the OLS estimation

[4] The selection of the neighborhoods can be done on a variety of criteria (administrative-type boundaries like postal codes and census tracts, or geographical boundaries like rivers, roads, andrail-tracks) that do not provide a unique answer.

of the hedonic model. The number of mixtures identified will correspond with the number of the different submarkets the housing market is segmented into. In what follows, the Finite Mixture Model is presented along with the standard approach used for its estimation (the Expectation Maximization EM algorithm). Subsequently, the Minimum Description Length criterion variant proposed by Figuereido and Jain (2002) will be presented and its relative merits over EM approaches will be discussed. In the folllowing sections, the random variables Y used in the exposition of the model will correspond to the residuals obtained from the hedonic model, each residual representing a different dwelling.

2.2.1 The Finite Mixture Model and the EM algorithm

Let $\mathbf{Y} = [Y_1, \ldots, Y_d]^T$ be a d-dimensional random variable, with $\mathbf{y} = [y_1, \ldots, y_d]^T$ representing one particular outcome of $\mathbf{Y}$. It is said that $\mathbf{Y}$ follows a k-component finite mixture distribution if its probability function can be written as:

$$p\left(\mathbf{y}|\ \theta\right) = \sum_{m=1}^{k} \alpha_m p\left(\mathbf{y}|\ \theta_m\right) \tag{2}$$

where $\alpha_1, \ldots, \alpha_k$ are the *mixing probabilities*, θ_m corresponds with the parameters defining the m^{th} distributional component, and $\theta \equiv \{\theta_1, \ldots, \theta_k, \alpha_1, \ldots, \alpha_k\}$ is the complete set of parameters needed to define the mixture. The mixing probabilities α_m satisfy the following conditions:

$$\alpha_m \geq 0,\ m = 1, \ldots, k,\ and\ \sum_{m}^{k} \alpha_m = 1 \tag{3}$$

Given the set $\mathcal{Y} = \left\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\right\}$ of n independent and identically distributed samples, the log-likelihood corresponding to a k-component mixture is

$$\log p\left(\mathcal{Y}|\ \theta\right) = \log \prod_{i=1}^{n} p(\mathbf{y}^{(i)}|\theta) = \sum_{i=1}^{n} \log \sum_{m=1}^{k} \alpha_m p(\mathbf{y}^{(i)}|\theta_m) \tag{4}$$

The *Maximum likelihood* (ML) estimate can not be found analytically. The usual choice for estimating the parameters of the mixture model is the employment of the *Expectation Maximization* (EM) algorithm. EM is an iterative procedure that finds the local maxima of the $\log p\left(\mathcal{Y}|\ \theta\right)$ and is extensively discussed in the work of McLachlan and Peel (2000) for the case of finite mixtures.

The EM algorithm is based on the interpretation that the data are incomplete. In the case of finite mixtures, the missing part is a set of n labels $\mathcal{Z} = \left\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(n)}\right\}$ associated with the n samples, indicating which component produced each sample. Each label is a binary vector $\mathbf{z}^{(i)} = [z_1^{(i)}, \ldots, z_k^{(i)}]$, where $z_m^{(i)} = 1$ and $z_p^{(i)} = 0$ for $p \neq m$, which means that the sample $\mathbf{y}^{(i)}$ was produced by the m^{th} component. Consequently, the complete log-likelihood can be written as

$$\log p\left(\mathcal{Y}, \mathcal{Z}|\ \theta\right) = \sum_{i=1}^{n} \sum_{m=1}^{k} z_m^{(i)} \log \alpha_m p(\mathbf{y}^{(i)}|\theta_m) \tag{5}$$

The EM algorithm results in a sequence of estimates $\left\{\widehat{\theta}(t),\ t = 0, 1, 2, \ldots\right\}$ by iteratively applying 2 alternate steps:

- **E-step**: Computes the conditional expectation of the complete log-likelihood, given $\mathcal{Y}$ and the current estimate $\widehat{\theta}(t)$. Since $\log p\left(\mathcal{Y},\mathcal{Z}|\ \theta\right)$ is linear with respect to the missing data $\mathcal{Z}$, we have to compute the conditional expectation $\mathcal{W} = E[\mathcal{Z} \mid \mathcal{Y}, \widehat{\theta}(t)]$ and then plug it back into the complete log-likelihood. The result is the so-called Q-function:

  $$Q(\theta, \widehat{\theta}(t)) \equiv E\left[\log p\left(\mathcal{Y},\mathcal{Z}|\ \theta\right) \mid \mathcal{Y}, \widehat{\theta}(t)\right] = \log p\left(\mathcal{Y},\mathcal{W}|\ \theta\right) \tag{6}$$

  Since the elements of $\mathcal{Z}$ are binary, their conditional expectations are given by

  $$\begin{aligned} w_m^{(i)} \equiv E\left[z_m^{(i)}|\mathcal{Y}, \widehat{\theta}(t)\right] &= \Pr\left[z_m^{(i)} = 1|\mathbf{y}^{(i)}, \widehat{\theta}(t)\right] && (7) \\ &= \frac{\widehat{\alpha}_m(t)\ p(\mathbf{y}^{(i)}|\widehat{\theta}_m(t))}{\sum_{j=1}^{k} \widehat{\alpha}_j(t)\ p(\mathbf{y}^{(i)}|\widehat{\theta}_j(t))} && (8) \end{aligned}$$

- **M-step:** Updates the parameter estimates according to

  $$\widehat{\theta}(t+1) = \arg\max_{\theta} \left\{Q(\theta, \widehat{\theta}(t)) + \log p(\theta)\right\} \tag{9}$$

  under the constraints of (3).

The problem with working with the EM algorithm is that it assumes that the number of mixture components k is known. This is never the case in real applications. The employment of the ML criterion is not appropriate; the class of the k-component models ($\mathcal{M}_k$) is nested in the class of the (k+1) components model ($\mathcal{M}_{k+1}$), and thus the likelihood will be increasing even if an extra, unnecessary mixture component is added. Alternative methods have been proposed in order to estimate the number of the mixture distributions. The most common ones are of a deterministic nature. A set of candidate models is estimated using EM for a range of values of k ($k \in [k_{min}, \ldots, k_{max}]$) which is assumed to contain the true value of k. The number of components can then be determined according to

$$\widehat{k} = \arg\min_{k} \left\{C\left(\widehat{\theta}(k), k\right), k = k_{min}, \ldots, k_{max}\right\} \tag{10}$$

where $C\left(\widehat{\theta}(k), k\right)$ is some selection criterion and $\widehat{\theta}(k)$ is the estimate of the set of the mixture parameters for the given k. The usual form of these criteria is:

$$C\left(\widehat{\theta}(k), k\right) = -\log p\left(\mathcal{Y}|\ \widehat{\theta}(k)\right) + \mathcal{P}(k) \tag{11}$$

where $\mathcal{P}(k)$ is an increasing function penalizing higher values of k. Such criteria are Approximate Bayesian criteria (*Laplace-empirical criterion* LEC, *Bayesian inference criterion* BIC), approaches based on information/coding theory concepts (*Akaike's information criterion* AIC, *Minimum message length criterion* MML), and methods based on the complete likelihood (5) (*Classification likelihood criterion* CLC, *Integrated classification likelihood criterion* ICL). For a more detailed review and comparison of these methods, the reader is directed to McLachlan and Peel (2000).

2.2.2 The Figuereido and Jain (2002) Minimum Message Length Criterion variant

The deterministic methods discussed in the previous section select a model class $\mathcal{M}_k$ based on its "best" representative $\widehat{\theta}(k)$. Nevertheless, in the mixture models, the distinction between model-class selection and model estimation is unclear, e.g. a 3-component mixture with one mixing probability equal to zero can not be distinguished from a 2-component mixture. Moreover, EM algorithms are known to suffer from two major drawbacks: (a) they are highly dependent on initialization values, and (b) they may converge to the boundary of the parameter space. Potential methods to correct for (a), include among others repeated applications of the EM algorithm for different initialization schemes. Nevertheless, these approaches demand a lot of computation time.

Figuereido and Jain (2002) proposed a different approach in order to circumvent these issues. They proposed a *Minimum Message Length* (MML) criterion that endogeneizes the selection of the optimal k. This is achieved by attempting to identify the "best" overall model in the entire set of available models

$$\bigcup_{k_{min}}^{k_{max}} \mathcal{M}_k$$

rather than selecting one among a set of candidate models.

The main idea behind minimum encoding length criteria (like MML and MDL) is that a model is a good one if it can be described by using a short code for the data. Consider some dataset $\mathcal{Y}$ that it is known to have been generated according to the probability model $p\left(\mathcal{Y}|\,\theta\right)$. Our goal is to encode the data and transmit them without any loss of information. If $p\left(\mathcal{Y}|\,\theta\right)$ is known to both the transmitter and the receiver, then they can both build the same code and communicate. However, if the parameters θ are unknown, the transmitter first needs to estimate them and subsequently transmit them. This leads to a two-part message, whose total length is given by

$$Length\left(\theta,\mathcal{Y}\right) = Length\left(\theta\right) + Length\left(\mathcal{Y}|\,\theta\right) \tag{12}$$

All minimum encoding length criteria state that the parameter estimate is the one that minimizes$Length\left(\theta,\mathcal{Y}\right)$. Figuereido and Jain (2002) obtained the following form of the MML criterion (derived in their Appendix), where the minimization with respect to θ is simultaneous in both θ (the parameter space) and its dimension c:

$$\widehat{\theta} = \arg\min_{\theta}\left\{-\log p\left(\theta\right) - \log p\left(\mathcal{Y}|\,\theta\right) + \frac{1}{2}\log\left|\mathbf{I}\left(\theta\right)\right| + \frac{c}{2}\left(1+\log\frac{1}{12}\right)\right\} \tag{13}$$

where $\mathbf{I}\left(\theta\right) \equiv -E\left[D_{\theta}^{2}\log p\left(\mathcal{Y}|\,\theta\right)\right]$ is the (expected) Fisher information matrix and $\left|\mathbf{I}\left(\theta\right)\right|$ denotes its determinant[5].

The MDL criterion can be obtained as an approximation to (13) and is given by

$$\widehat{c}_{MDL} = \arg\min_{c}\left\{-\log p\left(\mathcal{Y}|\,\widehat{\theta}\left(c\right)\right) + \frac{c}{2}\log n\right\} \tag{14}$$

whose two-part code interpretation is clear: $-\log p\left(\mathcal{Y}|\,\widehat{\theta}\left(c\right)\right)$ is the data code-length, while each of the c parameter set components requires a code-length proportional to $(1/2)\log n$. Nevertheless, $\mathbf{I}\left(\theta\right)$ is not

[5] D_{θ}^{2} denotes the Hessian matrix.

available, in general, in analytical form for the case of mixtures. In order to circumvent this problem, we replace it by the complete-data Fisher information matrix $\mathbf{I}_c\left(\theta\right) \equiv -E\left[D_\theta^2 \log p\left(\mathcal{Y}, \mathcal{Z} \middle|\, \theta\right)\right]$, which is an upper bound for $\mathbf{I}\left(\theta\right)$. The authors adopt a prior expressing lack of knowledge about the mixing parameters. In more detail

1. The parameters of the different components are assumed to be *a priori* independent and also independent of the mixing probabilities:

$$p\left(\theta\right) = p\left(\alpha_1, \ldots, \alpha_k\right) \prod_{m=1}^{k} p\left(\theta_m\right)$$

2. For each factor $p\left(\theta_m\right)$ and $p\left(\alpha_1, \ldots, \alpha_k\right)$, the standard noninformative Jeffrey's prior is adopted

$$p\left(\theta_m\right) \propto \sqrt{\left|\mathbf{I}^{(1)}\left(\theta_m\right)\right|}$$

$$p\left(\alpha_1, \ldots, \alpha_k\right) \propto \sqrt{\left|\mathbf{M}\right|} = \left(\alpha_1 \alpha_2 \ldots \alpha_k\right)^{-1/2}$$

for $0 \leq \alpha_1, \alpha_2, \ldots, \alpha_k \leq 1$ and $\alpha_1 + \alpha_2 + \ldots + \alpha_k = 1$.

Given the above assumption and noticing that for a k-mixture the dimension of θ_m is $c = Nk + k$, where N is the number of the parameters for each of the mixture distributions, (13) becomes

$$\widehat{\theta} = \arg\min_{\theta} \mathcal{L}\left(\theta, \mathcal{Y}\right) \tag{15}$$

with

$$\mathcal{L}\left(\theta, \mathcal{Y}\right) = \frac{N}{2} \sum_{m=1}^{k} \log\left(\frac{n\alpha_m}{12}\right) + \frac{k}{2} \log\frac{n}{12} + \frac{k(k+1)}{2} - \log p\left(\mathcal{Y} \middle|\, \theta\right) \tag{16}$$

The following intuitive interpretation can be given to this criterion:

1. $-\log p\left(\mathcal{Y} \middle|\, \theta\right)$ is the code-length of the data.

2. The expected number of points generated by the m^{th} component is $n\alpha_m$; this can be thought of as the effective sample size for the estimation of θ_m. This gives rise to the term $(N/2)\log(n\alpha_m)$.

3. The α_m's are estimated from all n observations, giving rise to the term $(k/2)\log(n)$.

The objective function in (16) does not make sense if any of the α_m's is equal to zero (it becomes $-\infty$). Regardless, the only thing that we need to code and transmit are the mixing elements whose probability is nonzero. Defining by k_{nz} the number of the nonzero probability components, (16) becomes

$$\mathcal{L}\left(\theta, \mathcal{Y}\right) = \frac{N}{2} \sum_{m:\alpha_m>0} \log\left(\frac{n\alpha_m}{12}\right) + \frac{k_{nz}}{2} \log\frac{n}{12} + \frac{k_{nz}(k+1)}{2} - \log p\left(\mathcal{Y} \middle|\, \theta\right) \tag{17}$$

Figueiredo and Jain (2002) also provide the variant of the EM algorithm that is necessary to estimate the finite mixture using (17). In particular, for fixed k_{nz}, the E-step is

$$\widehat{\alpha}_m(t+1) = \frac{\max\left\{0, \left(\sum_{i=1}^n w_m^{(i)}\right) - \frac{N}{2}\right\}}{\sum_{j=1}^k \max\left\{0, \left(\sum_{i=1}^n w_j^{(i)}\right) - \frac{N}{2}\right\}}, \tag{18}$$

$$m = 1, 2, \ldots, k$$

whereas the M-step is

$$\widehat{\theta}(t+1) = \arg\max_{\theta} Q(\theta, \widehat{\theta}(t)),\ \ for\ m : \widehat{\alpha}_m(t+1) > 0 \tag{19}$$

In the case of univariate Gaussian mixture models, each of the mixing distributions is fully specified by its mean and variance, i.e. $\theta_m = \left(\mu_m, \sigma_m^2\right)$. The M-step of the EM algorithm is then summarized in the following two equations:

$$\widehat{\mu}_m(t+1) = \left(\sum_{i=1}^n w_m^{(i)}\right)^{-1} \sum_{i=1}^n y^{(i)} w_m^{(i)} \tag{20}$$

$$\widehat{\sigma}_m^2(t+1) = \left(\sum_{i=1}^n w_m^{(i)}\right)^{-1} \sum_{i=1}^n \left(y^{(i)} - \widehat{\mu}_m(t+1)\right)^2 w_m^{(i)} \tag{21}$$

In order to improve the speed of convergence and to avoid problems that EM presents, such as sensitivity to initial conditions, the authors use the Component-Wise Expectation Maximization Algorithm (CEM^2) proposed by Celeux et al. (1999). CEM^2 permits the parameters' estimation of one mixture distribution at each iteration. Celeux et al. (1999) prove the convergence properties of CEM^2 by using the fact that EM falls into a class of iteration methods called *Proximal Point Algorithms* (Chretien and Hero III 2000).

2.2.3 Recovering the location of the submarkets

It was stated earlier that the number of submarkets will be equal to the estimated number of the mixing distibutions. Given the estimates from the finite mixture model, it is necessary to provide a decision rule that will reconstruct the market segments. In order to do so, the *Maximum Posterior Mode* (MPM) algorithm will be used. The MPM algorithm assigns to each data point, here each dwelling, the mixing distribution that is the most probable to have generated the particular point value. More specifically, the data point *(i)* is allocated to the m^{th} distribution if

$$w_m^{(i)} > w_j^{(i)}, \quad \left(j = 1, 2, \ldots, \widehat{k}\right) \tag{22}$$

where $w_m^{(i)}$ comes from (7). The different submarkets are then defined as the collection of the housing data points assigned to each distribution.

3 Simulation Results

The proposed methodology for the identification of regions has been introduced in the context of engineering applications, such as image segmentation problems where the input data are pixel intensities. Thus, it is of interest to examine the performance of the method when economic data, such as housing prices, are considered. Two different types of simulations were performed in order to examine the ability of the model to identify housing submarkets. The first one involved the simple case of identifying regions defined by points generated as independent draws from different distributions. The second one attempted to imitate a genuine problem by considering a hypothetical housing market divided into different submarkets.

3.1 Case I: IID regions

This set of simulations investigated the performance of the estimation procedure for the finite mixture model alone. An two-dimensional area was divided into three different segments. Each segment contained points generated independently by a Normal distribution. I considered three different cases of increasing variability for the values of the points. The theoretical parameters for each distribution and their corresponding estimates are summarized in Tables 1 to 3.

The following criteria need to be satisfied for ensuring the validity of the proposed model. The first criterion concerns the correct estimation of the underlying distributions' parameters; the estimates obtained were quite accurate being more precise for low-variability data. The second criterion corresponds to the correct identification of the segments' boundaries; Figures 1 to 3 depict the identified segments as a collection of points in space with different colors. Points with the same color were identified to have been generated from the same distribution, i.e. belonging to the same submarket. The separation of the different colored segments is visible in the figures. Once more, the level of separation was more precise for the low and medium-variability data. The last criterion amounts to the degree of homogeneity within each identified segment; high level of homogeneity will be represented in the figures by high color homogeneity as well. The degree of homogeneity is excellent for the low-variability data (Figure 1), where the segments do not contain points that are attributed to a different segment (100% homogeneous segments) ,very good for the medium-variability data (Figure 2), where the segments do contain a few points that are attributed to a different segment (approximately 99% homogeneous segments), and good for the high-variability data (Figure 3), where the segments are, on average, approximately 10% homogeneous. Thus, misclassification of points to a different that their true segment may occur at most at 10% of the time, which can be considered acceptable.

Overall, the performance of the model in terms of the aforementioned criteria was good, especially for the cases of the low and the medium variance of the underlying distributions.These results validated the model's adequacy in estimating finite mixture model for at least the simple case of IID regions.

3.2 Case II: Submarkets

The case of IID regions is not very realistic. Hence the ability of the proposed model should be tested in cases that emulate an actual housing market as much as possible. In order to do so, a hypothetical housing market segmented into seven different submarkets (see Figure 4) was constructed. For each submarket, a hedonic model based on a set of structural characteristics was assumed. The structural characteristics' values were created using a variety of distributional configurations. The details of this procedure are presented in section 2.6.

Preliminary results obtained from the estimation on the Los Angeles data were used for calibrating the simulation parameters. Most of the structural characteristics' values were generated independently from the values of other characteristics. The only exceptions were the values for the number of bathrooms (the correlation with the number of bedrooms was about 0.85), the values for the structure area (the correlation with the number of bedrooms was about 0.87), and whether the dwelling would have a garage or not (the correlation with the existence of a fireplace was about 0.69). The following functional form that was assumed in order to generate the dwelling log prices for each of the assumed seven submarkets is[6]:

$$\begin{aligned} log(SV) = \alpha + \beta_1 lot + \beta_2 structure + \beta_3 age + \beta_4 bed + \beta_5 bath/bed \\ + \beta_6 fireplace + \beta_7 garages + \beta_8 pool + \beta_9 owner + \beta_{10} Xcor + \beta_{11} \\ + Xcor^2 + \beta_{12} Ycor + \beta_{13} Ycor^2 + \epsilon \end{aligned} \tag{23}$$

The simulation experiment was conducted for two alternative error (ϵ) variance specifications. The first one assumed a high standard deviation of 0.5, which corresponded with the unrealistically difficult case in which the standard error of deviation from the true hedonic value is 50% of the price. The second specification assumed a lower standard deviation of 0.1, which corresponds with 10% of the house price as the standard error of the deviation.

The results of the simulations for the two cases are depicted in Figures 5 and 6 and tabulated in Tables 4 to 7. The evaluation of the results will be based on the same criteria as in the IID case; higher degree of color homogeneity is associated with higher segment homogeneity.

In particular, the following points can be noted. The estimation success is measured in terms of the adjusted R^2 of the regressions for each of the identified regions. These measures suggested that the model had similar if not higher overall explanatory power when compared to the values of the regressions for the true regions. Regarding the second criterion of the correct identification of the segments' boundaries, the boundaries of the different regions were identified, with the exception of the boundary between regions 2 and 3, as can be seen in the figures (there is no color discrepancy in the regions' boundary). In particular, the identification of the boundaries of regions 4 and 7, which are elliptical and thus the most complex, was very successful. The final criterion of within-region heterogeneity was evaluated, numerically, by the following measure: for each of the true regions, the percentage of points belonging to each of the identified regions was calculated. The results are tabulated in Table 5 in detail, and summarized in Table 6. A high value would indicate a high degree of homogeneity, and thus the true region would be correctly identified. Regions 1 and 7 were "strongly" identified (the maximum proportion rates were above 85% for both variability specifications), regions 4 and 5 were "identified" (the maximum proportion rates were above 45% and double in size from the next rate corresponding to another identified region), whereas the other regions were not "identified" and, as can also be visually verified by the figures, had a large degree of within heterogeneity. These results are also evident by the color homogeneity within-regions; more colors (submarkets) are appearing in regions 3 and 6. The effect seems to be more pronounced for the case of low error variance compared with the high error variance, which is something that it is not expected especially in view of the results for the IID regions case.

The large degree of the heterogeneity and the unexpected higher variation in the identified submarkets for the low error variance case may be explained by the choice of the way the submarkets are recovered.

[6] The table with the corresponding parameter values used to construct the log prices is provided in section 2.6 for each one of the seven submarkets.

The MPM approach, treats each point individually without taking into consideration the neighboring points, and although this criterion maximizes the expected number of correct states, there could be some problems with the resulting identified submarkets (the "optimal" segmentation may, in fact, not even be a valid segmentation). This occurs because MPM determines the most likely submarket for each dwellings separately, without regard to the probability of groups of dwellings . In order to circumvent this problem, one may attempt to maximize the expected numbers of correct assignment to a submarket of pairs of dwellings, or even of triples of dwellings . The extreme version of this approach, which is the Maximum A Posteriori (MAP) approach, is to find the single best segmentation for all data points (dwellings) together. This is accomplished by the Viterbi algorithm, which is an iterative procedure, and thus requiring a large number of extra calculations that is prohibiting for the case of larger datasets.

Regardless, even with the employment of the MPM criterion for the recovery of the submarkets, the model seemed to identify adequately most of the imposed submarkets and to explain most of the variation in sales prices, and thus it can be considered suitable for use in empirical applications.

4 Los Angeles Data

4.1 Data description

The model was applied to the market values of dwellings sold in the Los Angeles county during 2002, with a map of the area presented in Figure 7. These data are provided by DataQuick, which repackages information purchased by the Los Angeles county Assessor's office. The dataset provides a variety of characteristics for each dwelling in the area of Los Angeles along with its assessed value and the market value for the most recent transaction. The analysis was focused in the market values (measured in dollars) of single residences. I considered a semi-logarithmic hedonic model. The chosen structural characteristics entering as independent variables were the size of the lot (in ft^2), the size of the structure (in ft^2), the age of the dwelling (in years), the number of bedrooms, the number of bathrooms, the number of garages, the existence of a fireplace, the existence of a pool and whether the dwelling was inhabited by its owner or not. For the hedonic model specification, the ratio of the number of bathrooms over the number of bedrooms was selected instead of just the number of bathrooms for two reasons; the number of bathrooms is highly correlated with the number of bedrooms and the presence of both in the regression may cause near-collinearity problems, and we consider it to be a better measure of the comfort/utility of the house. The geographical coordinates for each dwelling were obtained by Geocoding performed in ArcGIS software.

The data were not completely unproblematic. Missing values for the structural characteristics that could be attributed to problems in data collection necessitated some preliminary filtering of the data. Dwellings for which values for a characteristic were missing were removed from the sample. Moreover, geocoding did not provide geographical coordinates for all dwellings[7]. Thus, a final sample of 108488 dwellings, sold in the year 2002, were available out of a total sample of approximately 2000000 residences. The summary statistics and the correlation matrix of the market values and the structural characteristics are provided in Table 9. The following features can be observed:

1. The number of bedrooms, the number of bathrooms and the area of the dwelling structure were highly

[7] This is not unusual during the application of geocoding, as discussed by Reichert 2002.

correlated (>85%), as would be anticipated. On the other hand, the large correlation (60%) between the existence of fireplace and the number of garages might not be so obvious.

2. If the medians are considered, the typical dwelling cost \$250000, had 3 bedrooms and 2 bathrooms, had a garage, its structure area was 1800 ft^2 located in a lot of 7600 ft^2, it was constructed 45 years ago, and it was inhabited by its owner.

4.2 Results

A semi-logarithmic hedonic model was estimated according to (23) by Ordinary Least Squares. The obtained residuals were then used in order to estimate the number and the location of the different submarkets [8]. The algorithm resulted in identifying four different submarkets. The small number of identified submarkets suggests that the attempt to determine them by using administrative or geographical boundaries might have been misleading. The Los Angeles county is divided into 88 cities. Thus, where there may appear to be more diverse submarkets based on administrative boundaries or other considerations, the statistics identify only a small number. This suggests that neighborhood characteristics and other omitted variables can be captured more parsimoniously by only a few submarkets with different hedonic parameters. Things are simpler than might appear in a first glance.

The identified submarkets are depicted in Figure 8 (each color corresponds to a different submarket) and the mean values of the structural characteristics are tabulated in Table 10. The sizes of the submarkets differed substantially. Identified submarket 2 contained more than 80% of the total dwellings and it covered most of the Los Angeles County area. Submarket 3, which was the second largest and accounted for more than 16% of total dwellings number, also contained points in the entire Los Angeles county. The points that belonged to the other two submarkets (1,4) presented no particular geographical pattern even though there seems to be a concentration of dwellings of submarket 1 with dwellings belonging to submarket 3.

Even though, there seems not to exist a closed homogenous regions corresponding to particular submarkets, there do exist locations in which there is a high concentration of dwellings that belong to the submarkets 1 and 3. In particular, when Figure 8 is compared with Figure 7 (the map of Los Angeles county), it can be noticed that such dwellings are found to be concentrated in the areas of Malibu, Beverly Hills and West Hollywood, Rancho Palos Verdes, Rolling Hill Estates, Long Beach, Pomona and Pasadena. These results are not a surprise; submarkets 1 and 3 contain houses with higher values that may be attributed to paying a location premium. The mean dwelling values are around \$635.000 for submarket 3 and \$782.000 for submarket 1. Thus, a pattern seems to emerge in geographical terms for the case of submarket 3, since dwellings belonging to this submarket exists in high concentration on premium locations in the Los Angeles county.

Moreover, a closer examination of Table 10 suggests that the existence of the submarkets may be also be attributed to differences in the structural characteristics of the dwellings besides their prices. For example, differences in the mean number of bedrooms, the average lot size, and the average structure area exist among the submarkets. Thus, the identification of different submarkets might be a result of the disparities in the implicit prices of the structural characteristics across diverse types of dwellings.

The next step in the model was the estimation of the hedonic index for each of the identified submarkets in

[8] The model was programed in R software and the estimation was done by using a Power Mac G5 dual processor 2.0 GHz with 4 GB of RAM. The $MDL - CEM^2$ estimation took approximately 24 hours to be completed.

order to determine the implicit prices for the structural characteristics. The results along with the regression results for the entire sample are tabulated in Table 11. We note that:

- In all regressions but the one corresponding to the 4^{th} identified region, which is also the smallest of all with only 963 dwellings (<1% of total dwellings), most of the parameter estimates were statistically significant and of the expected sign; age contributed negatively to the value of the house, whereas the size of the lot, the size of the structure, the existence of fireplace, garages contributed positively. The coefficient for the ratio of the number of bathrooms over the number of bedrooms was positive indicating that more bathrooms contribute positive to the dwelling value.

- The coefficient for the number of bedrooms was found negative. Although more bedrooms might seem to have an unambiguously positive effect, if they are considered to represent, for a given living area, the division of the floor area, then an explanation for this effect can be provided. Much of the housing stock was developed when families were larger (the median age of dwellings in the sample was 47 years), and thus there was a demand for more bedrooms. Nowadays, the size of the family has reduced and married couples with no children might want a lot of open area; more bedrooms would replace the amount of open area given a particular living area size, and thus a lot of bedrooms are no longer desirable and thus there is a reduced demand for dwellings with many bedrooms. Since the hedonic model represents the equilibrium point between demand and supply in the real estate market, this reduced demand can explain the negative effect for an extra bedroom, when the total living area is accounted for.

- The hedonic models for each submarket have higher R^2 compared to the one obtained for the entire market. In particular, R^2 for the regression estimated in the second submarket, which amounted to 79.2% of the total market, was equal to 0.5811, whereas R^2 for the regression estimated in the third submarket, which amounted to 16.6% of the total market, was even higher and equal to 0.8126.

- The F-test confirmed the superiority of the submarkets model over a model that assumed no spatial heterogeneity with a p-value equal to zero.

- The economic significance of the estimated coefficients for the structural characteristics varied across submarkets[9]. The impact of an extra 1000 ft^2 for the structure area was an 15% increase in the value of the dwelling and it was rather stable across submarkets. The impact of extra 1000 ft^2 for the lot area is quite smaller and almost negligible in economic terms leading to an average increase of the value of about 0.42%. Similar values are also observed in the literature: Palmquist (1992), Goodman and Thibodeau (1995). On the other hand, the impact of the existence of a fireplace accounted for an increase in the house price of 54%, 19%, and 22.6% for submarkets 1, 2 and 3 respectively. A similar pattern of decaying impact on the value across submarkets was apparent for the existence of the pool (32%,17%, and 10%), and whether the dwelling is inhabited by its owner.

- In dollar terms, the effect of changes of values in the structural characteristics will depend on the average house values for each identified submarket. The mean dwelling prices are approximately \$780000, \$270000, and \$630000 for the first, second, and third submarket respectively. Thus, the

[9] The analysis is concentrated in the first 3 submarkets that amount for more than 99% of the total market.

existence of a fireplace will amount to an increase in the dwelling value by $421000, $51000, and $142000 for the each three submarkets respectively, whereas the existence of a pool will increase the value by $249000, $46000, and $62000.

Consequently, the proposed approach that implicitly models spatial heterogeneity offers an improvement upon a simple hedonic model by identifying, in a very parsimonious way, the underlying housing submarkets. In particular, the statistical identification of only four submarkets versus the administrative division of the Los Angeles county to 88 cities (and even more communities) can be used in order to simplify the mass assessment processes. In practice, there can be considered only two different relations that can assess dwellings value that may be based on the identified submarkets no 2 and 3 that account for about 96% of all single family houses in the county. The first relation will be used in order to assess values in high premium areas such as Malibu, Beverly Hills, Long Beach etc, and it will be based on the estimated coefficients of the hedonic specification for submarket 3. The remaining dwellings in the county will be assessed using the coefficients for submarket 2. The enormous amount of simplification in using the identified submarkets versus the administratively imposed ones is obvious.

Given that the main focus was the determination of the statistically identified submarkets, no comparison in terms of the accuracy was provided as such a comparison would be heavily depended on the choice of the type of administrative boundaries.

5 Conclusion

The approach described here emphasizes classification accounting for heterogeneity, by making groups of observations define potential submarkets. In hedonic studies, submarkets are typically defined a priori. I provide a systematic way to identify separate submarkets, and to decompose differences among the submarkets into price and quantity/quality effects. This method allows the researcher to model rather than impose submarkets and allows for the consideration of large datasets. In particular, a mixture model was applied to let the data determine the market structure and subsequently the hedonic model is re-estimated for each submarket. The estimation ability of the proposed model was validated through two different sets of simulations.

The model was subsequently applied to prices of Los Angeles county dwellings that were sold during the year 2002. The findings suggested that the city's housing market could be segmented into only 4 different submarkets with different hedonic specifications compared to a total of 88 submarkets defined by cities' limits. The parameter estimates of the hedonic models for each of the segments were significant both in statistical and economical terms and of the correct sign. Dwellings belonging to one of the identified submarkets are found to be clustered in high premium residential areas in the county such as Malibu, Beverly Hills, Long Beach, etc. Additionally, the identified submarkets can be partly explained by differences in the structural characteristics of dwellings. This points to the observation that the variation attributed to neighborhood characteristics can be captured by a rather limited number of distinct submarkets of diverse types of dwellings.

Overall, the model managed to identify housing submarkets in a more parsimonious way than would have resulted from assuming the existence of submarkets solely in terms of geographicaly or administratively determined regions, thus making mass assessment of dwelling prices considerably simpler and faster.

References

[1] Adair A.S., J.N. Berry, and W.S . McGreal, (1996). Hedonic modelling, housing submarkets and residential valuation, *Journal of Property Research*, 13, 67-83.

[2] Adelman I. and Z. Griliches, (1961). On An Index of Quality Change, *Journal of the American Statistical Association*, 56, 535-548.

[3] Anselin L. (1988). *Spatial Econometrics: Methods and Models*, Dordrecht, Netherlands: Kluwer Academic.

[4] Basu S. and T. Thibodeau, (1998). Analysis of Spatial Autocorrelation in House Prices,*Journal of Real Estate Finance and Economics*, 17, 61-86.

[5] Can A., (1992). Specification and Estimation of Hedonic Housing Price Models, *Regional Science and Urban Economics*, 22, 453-474.

[6] Can A. and I.F. Megbolubge, (1997). Spatial Dependence and House Price Index Construction, *Journal of Real Estate Finance and Economics*, 14, 203-222.

[7] Celeux G., S. Chrétien, F. Forbes, and A.Mkhadri, (1999). A Component-Wise EM Algorithm for Mixtures, *Technical Report 3746, INRIA Rhne-Alpes, France.*

[8] Chrétien S. and A.O. Hero III, (2000). Kullback Proximal Algorithms for Maximum-Likelihood Estimation, *IEEE Transactions on Information Theory*, 46, 1800-1810.

[9] Clapp J.M., H-J. Kim, and A.E. Gelfand, (2002). Spatial Prediction of House Prices Using LPR and Bayesian Smoothing, *Real Estate Economics*, 30, 505-532.

[10] Clapp J.M. and Y.Wang, (2005). Defining Neighborhood Boundaries: Are Census Tracts Obsolete?,*Working paper*

[11] Dubin R., (1992). Spatial Autocorrelation and Neighborhood Quality, *Regional Science and Urban Economics*, 22, 433-452.

[12] Dubin R., (1998). Predicting House Prices Using Multiple Listing Data, *Journal of Real Estate Finance and Economics*, 17, 35-60.

[13] Dubin R., Pace R.K., and T.Thibodeau, (1999). Spatial Autoregression Techniques for Real Estate Data, *Journal of Real Estate Literature*, 7, 79-95.

[14] Epple D., (1987). Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Product, *Journal of Political Economy*, 95, 59-80.

[15] Figuereido M.A.T. and A.K. Jain, (2002). Unsupervised Learning of Finite Mixture Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 381-396.

[16] Freeman III A.M, (1979). Hedonic Prices, Property Values and Measuring Environmental Benefits, *Scandinavian Journal of Economics*, 81, 154-173.

[17] Goodman A.C., (1981). Housing Submarkets within Urban Areas: Definitions and Evidence, *Journal of Regional Science*, 21, 175-185.

[18] Goodman A.C. and T. Thibodeau, (1998). Housing Market Segmentation, *Journal of Housing Economics*, 7, 121-143.

[19] Kain J.F and J.M. Quigley, (1970). Measuring the Value of Housing Quality, *Journal of the American Statistical Association*, 65, 532-548.

[20] McLachlan C. and D.Peel, (2000). *Finite Mixture Models*, New York: John Wiley and Sons.

[21] Mark J. and M.A. Goldberg, (1988). Multiple Regression Analysis and Mass Assesment: A review of issues, *The Appraisal Journal*, 56, 89-109.

[22] Meese R. and N.Wallace, (1997). The Construction of Residential Housing Pricing Indices: A Comparison of Repeat-Sales, Hedonic-Regression, and Hybrid Approaches, *Journal of Real Estate Finance and Economics*, 14, 51-73.

[23] Pace R.K, C.F. Sirmans, and V.C.Slawson Jr. (2002). Automated Valuation Models, *Real Estate Valuation Theory*, Boston: Kluwer Academic.

[24] Palmquist R.B., (1992).Valuing Localized Externalities, *Journal of Urban Economics*, 31, 59-68.

[25] Reichert A.K., (2002). Hedonic Modeling in Real Estate Appraisal: The case of Environmental Damages Assesment, *Real Estate Valuation Theory*, Boston: Kluwer Academic.

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

[27] Straszheim M., (1974). Hedonic Estimation of Housing Market Prices: A Further Comment, *The Review of Economics and Statistics*, 56, 404-406.

[28] Straszheim M., (1975). *An econometric analysis of the urban housing market*, New York : National Bureau of Economic Research, Columbia University Press.

[29] Tobler W., (1970). A computer movie simulating urban growth in the Detroit region, *Economic Geography*, 46, 234-240.

[30] Ugarte M.D., T. Goicoa, and A.F.Militino, (2004). Searching for Housing Submarkets Using Mixtures of Linear Models,*Advances in Econometrics*, 18, 259-276.

A Submarkets simulation study

A.1 Simulated structural characteristics

Str. Characteristic	Unit of measurement
Age	years
Lot	(ft^2)
Structure	(ft^2)
Bed	no of bedrooms
Bath/Bed	no of baths/no of beds
Fireplace	1/0
Pool	1/0
Owner	1/0
Garages	1/0
Xcor	longitude coordinate (projected)
Ycor	latitude coordinate (projected)

A.2 Generating the variables

- **Number of bedrooms**: the values for this variable were obtained by random draws from a normal distribution rounded up to the nearest integer value. The mean and the variance for each of the submarkets are given below.

Segment	mean	variance
1	4.3	50.5
2	3	5.3
3	3.8	6.8
4	6.8	213
5	5	7
6	6.1	180
7	2.8	5.1

- **Number of bathrooms**: the values for this variable were obtained by random draws from a normal distribution rounded up to the nearest integer. The mean was the number of bedrooms for each location point and the variance was constant and equal to unity.

- **Age**: the values for the Age variable were obtained by random draws from a mixture of three normal distributions rounded down to the nearest integer. The parameters for each of the submarkets are given below.

Segment	m1	v1	mixpr1	m2	v2	mixpr2	m3	v3	mixpr3
1	20	70	0.31	50	75	0.41	80	90	0.28
2	20	57	0.36	50	56	0.5	80	70	0.14
3	20	68	0.34	50	73	0.41	80	63	0.25
4	20	60	0.29	50	62	0.48	80	142	0.23
5	20	70	0.3	50	70	0.4	80	85	0.3
6	20	62	0.33	50	60	0.46	80	110	0.21
7	20	55	0.35	50	52	0.49	80	60	0.16

- **Pool**: the variable Pool (whether the dwelling has a pool or not) was constructed by random draws from a uniform distribution in [0,1] domain. Values for the entire population were created, but pool was assigned to the dwelling whose value was below the threshold value given in the table below (the value corresponds to the proportion of population with a swimming pool).

Segment:	1	2	3	4	5	6	7
Threshold Value:	0.1	0.11	0.12	0.14	0.13	0.15	0.06

- **Owner**: the variable Owner (whether the owner resides in the dwelling or not) was constructed in an analogous to the Pool variable way. The threshold values are tabulated in

Segment:	1	2	3	4	5	6	7
Threshold Value:	0.6	0.89	0.79	0.48	0.45	0.55	0.65

- **Fireplace**: the variable Fireplace (whether the dwelling has a fireplace or not) was constructed in an analogous way that the Pool and Owner variables were created. The threshold values are tabulated in

Segment:	1	2	3	4	5	6	7
Threshold Value:	0.27	0.35	0.35	0.28	0.3	0.4	0.25

- **Garages**: the variable Garages (whether the dwelling has a garage or not) was created by random draws from the Fireplace variable[10]. The proportion of the Fireplace population that was also assigned a garage is given in the following table:

Segment:	1	2	3	4	5	6	7
Threshold Value:	0.68	0.58	0.67	0.58	0.65	0.55	0.7

[10] Our preliminary data analysis indicated that the dwellings that have a garage are a proportion of the dwellings that also have a fireplace.

- **lot**: the values for the variable lot (the surface of the dwelling's lot in ft^2) were obtained by random draws from a normal distribution whose mean and variance, for each of the submarkets, are given below.

Segment	Mean	Variance
1	11000	9000000
2	6800	4000000
3	8500	6250000
4	15000	25000000
5	10000	8400000
6	14000	25000000
7	6000	3240000

- **Structure**: the variable structure (the surface of the dwelling's structural area in ft^2) is constructed by a linear combination of the number of bedrooms, the number of bathrooms, and the lot area plus a constant term. The corresponding parameters are tabulated below.

Segment	Constant	No of Bedrooms	No of Bathrooms	Lot Area
1	446.3	123.8	485.8	0.0033
2	57.3	211.2	371.9	0.0089
3	486	168.8	374.1	0.00179
4	207.5	97.2	637.3	0.00605
5	400	120	450	0.004
6	200	92	610	0.0063
7	58	200	380	0.01

- **Xcor**: the variable Xcor (the longitude coordinate) was originally set to take values from 1 to 500 (each integer value to correspond to one unit of longitude). In order to be conformable with real-world coordinate values, it was transformed to take values from (-119 (=0) to -117.5 (=500)).

- **Ycor**: the variable Ycor (the latitude coordinate) was originally set to take values from 1 to 500 (each integer value to correspond to one unit of latitude). In order to be conformable with real-world coordinate values, it was transformed to take values from (33.3 (=0) to 34.5 (=300)).

- **Xcor2**: the square of the Xcor variable

- **Ycor2**: the square of the Ycor variable

- **Sales value**: the logarithm of the sales values for each dwelling constructed according to (23) in the main text. The parameter values for each submarket are provided in the following table.

Segment	**Constant**	**lot**	**structure**	**Age**	**Bed**	**Bath/Bed**	**Fireplace**
1	10000	0.000004	0.000200	0.0008	-0.04	-0.015	0.15
2	-4300	0.000005	0.000150	-0.0023	-0.033	0.009	0.21
3	-3000	0.000005	0.000150	-0.0009	-0.013	0.11	0.25
4	17000	0.000048	0.000029	-0.0015	0.035	0.15	0.2
5	10250	0.000006	0.000150	-0.001	0.028	0.1	0.18
6	15000	0.000055	0.000035	-0.002	0.04	0.13	0.22
7	-4500	0.000005	0.000170	-0.0028	-0.038	0.01	0.18
Segment	**Garages**	**Pool**	**Owner**	**Xcor**	**Xcor2**	**Ycor**	**Ycor2**
1	0.32	0.54	0.175	240	1.009	245.5	-3.52
2	0.17	0.19	0.1	-45	-0.195	98.1	-1.4
3	0.1	0.23	0.065	-18.48	-0.08	115	-1.7
4	0.14	0.39	-0.23	360.2	1.52	253.9	-3.7
5	0.35	0.5	0.16	242	1.008	245.5	-3.51
6	0.15	0.4	-0.2	341.1	1.5	251.2	-3.6
7	0.16	0.16	0.08	-48.88	-0.21	100	-1.5

B Tables

B.1 Simulation Results: IID case

Case I: Low Variance			
No of Distributions	3	No of Points	15382
Distribution index	Mean	Standard Deviation	Mixing Probability
1	240000	4000	0.50
2	200000	4000	0.25
3	280000	4000	0.25
Case I: Estimates			
1	240015.58	3993.14	0.4582
2	199962.23	4027.90	0.2676
3	280010.49	4066.32	0.2742

Table 1

Case II: Medium Variance			
No of Distributions	3	No of Points	15382
Distribution index	Mean	Standard Deviation	Mixing Probability
1	280000	10000	0.50
2	200000	10000	0.25
3	240000	10000	0.25
Case II: Estimates			
1	279532.86	10524.22	0.2816
2	200215.38	10316.60	0.2706
3	239871.75	9537.06	0.4478

Table 2

Case III: High Variance			
No of Distributions	3	No of Points	15382
Distribution index	Mean	Standard Deviation	Mixing Probability
1	240000	20000	0.50
2	280000	20000	0.25
3	200000	20000	0.25
Case III: Estimates			
1	236801.03	16331.80	0.3705
2	275602.01	21342.55	0.3558
3	199105.13	19204.50	0.2737

Table 3

B.2 Simulation Results: Submarkets case

				Counts				
Ident/ True	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**no obs**
1	5	3515	7111	691	19765	4302	11	35400
2	540	120	6154	92	4153	2984	12	14055
3		2035	1714	1650	2594	3252	24	11269
4	200		66	10		6	2507	2789
5	6296	4	10227	308	692	3519	98	21144
6	24	216	2072	41	3327	1050	4	6734
7			8	102		119		229
no obs	7065	5890	27352	2894	30531	15232	2656	91620
				Frequencies				
Ident/True	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**total**
1	0.07%	59.68%	26.00%	23.88%	64.74%	28.24%	0.41%	38.64%
2	7.64%	2.04%	22.50%	3.18%	13.60%	19.59%	0.45%	15.34%
3	0.00%	34.55%	6.27%	57.01%	8.50%	21.35%	0.90%	12.30%
4	2.83%	0.00%	0.24%	0.35%	0.00%	0.04%	94.39%	3.04%
5	89.12%	0.07%	37.39%	10.64%	2.27%	23.10%	3.69%	23.08%
6	0.34%	3.67%	7.58%	1.42%	10.90%	6.89%	0.15%	7.35%
7	0.00%	0.00%	0.03%	3.52%	0.00%	0.78%	0.00%	0.25%
total	8%	6%	30%	3%	33%	17%	3%	100%

Table 4: High Error Variance

Counts

Ident/True	1	2	3	4	5	6	7	no obs
1	139		65	9		3	2524	2740
2			12	153	1	241	4	411
3		1299	3479	176	11948	1975	4	18881
4	18	293	4466	86	6480	2269	9	13621
5	828	24	6260	85	2519	3025	12	12753
6	6080		8492	294	127	2753	78	17824
7		1580	2247	282	6040	1355	5	11509
8		2694	2331	1809	3416	3611	20	13881
no obs	7065	5890	27352	2894	30531	15232	2656	91620

Frequencies

Ident/True	1	2	3	4	5	6	7	total
1	1.97%	0.00%	0.24%	0.31%	0.00%	0.02%	95.03%	2.99%
2	0.00%	0.00%	0.04%	5.29%	0.00%	1.58%	0.15%	0.45%
3	0.00%	22.05%	12.72%	6.08%	39.13%	12.97%	0.15%	20.61%
4	0.25%	4.97%	16.33%	2.97%	21.22%	14.90%	0.34%	14.87%
5	11.72%	0.41%	22.89%	2.94%	8.25%	19.86%	0.45%	13.92%
6	86.06%	0.00%	31.05%	10.16%	0.42%	18.07%	2.94%	19.45%
7	0.00%	26.83%	8.22%	9.74%	19.78%	8.90%	0.19%	12.56%
8	0.00%	45.74%	8.52%	62.51%	11.19%	23.71%	0.75%	15.15%
total	8%	6%	30%	3%	33%	17%	3%	100%

Table 5: Low Error Variance

	True Regions						
Max Proportion	**1**	**2**	**3**	**4**	**5**	**6**	**7**
High Error Variance	89%	60%	37%	57%	65%	28%	94%
Low Error Variance	86%	46%	31%	63%	39%	24%	95%

Table 6: Maximum concentration of identified regions points per true region

True Regions								
adjusted R2	**All obs**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
High Error Variance	**0.705**	**0.861**	**0.515**	**0.597**	**0.846**	**0.965**	**0.942**	**0.465**
Low Error Variance	**0.714**	**0.994**	**0.964**	**0.974**	**0.993**	**0.999**	**0.998**	**0.956**
no observations	**91620**	**7065**	**5890**	**27352**	**2894**	**30531**	**15232**	**2656**
Identified Regions								
adjusted R2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
High Error Variance	**0.984**	**0.986**	**0.976**	**0.910**	**0.923**	**0.999**	**0.713**	**na**
no observations	**35400**	**14055**	**11269**	**2789**	**21144**	**6734**	**229**	**na**
Low Error Variance	**0.941**	**0.762**	**0.996**	**0.994**	**0.983**	**0.941**	**0.998**	**0.977**
no observations	**2740**	**411**	**18881**	**13621**	**12753**	**17824**	**11509**	**13881**

Table 7: Region Regressions adjusted R^2

B.3 Data Filtering results

	Dwellings left		Dwellings left
No filtering	166212	**Structure**	128936
Value	137561	**Bedrooms**	123659
Age	133570	**Bathrooms**	123624
Lot	129381	**Geocoding**	108488

B.4 Real Data Results: Summary Statistics

Summary Statistics: Unsegmented Market										
	Values	**Age**	**Bathrooms**	**Bedrooms**	**Fireplace**	**Garages**	**Pool**	**Lot size**	**Structure Size**	**Owner**
Min.	100	0	1	1	0	0	0	4	114	0
Q1	186000	26	1	2	0	0	0	4230	1101	1
Median	262500	47	2	3	0	1	0	6075	1420	1
Mean	343600	45.68	2.438	3.305	0.3465	0.5092	0.1096	7591	1810	0.8633
Q3	394000	61	3	4	1	1	0	7680	1928	1
Max	10590000	199	124	124	1	4	1	857300	187000	1

Table 8: Summary Statistics

Correlation Matrix: Unsegmented market										
	Value	**age**	**bath**	**bed**	**fireplace**	**garages**	**pool**	**lot**	**structure**	**owner**
Value	1.000	0.003	0.422	0.381	0.195	0.085	0.240	0.181	0.490	-0.116
age	0.003	1.000	-0.117	-0.025	0.176	0.281	-0.046	-0.009	-0.042	-0.116
bath	0.422	-0.117	1.000	0.875	-0.073	-0.167	0.089	0.113	0.882	-0.260
bed	0.381	-0.025	0.875	1.000	-0.035	-0.086	0.093	0.124	0.836	-0.263
fireplace	0.195	0.176	-0.073	-0.035	1.000	0.600	0.265	0.077	-0.004	0.099
garages	0.085	0.281	-0.167	-0.086	0.600	1.000	0.184	0.044	-0.104	0.143
pool	0.240	-0.046	0.089	0.093	0.265	0.184	1.000	0.130	0.135	0.041
lot	0.181	-0.009	0.113	0.124	0.077	0.044	0.130	1.000	0.172	-0.040
structure	0.490	-0.042	0.882	0.836	-0.004	-0.104	0.135	0.172	1.000	-0.247
owner	-0.116	-0.116	-0.260	-0.263	0.099	0.143	0.041	-0.040	-0.247	1.000

Table 9: Correlation Matrix

B.5 Real Data Results: Identified regions

Mean Values: Segmented Market										
Ident. Region	**Values**	**Age**	**Bathrooms**	**Bedrooms**	**Fireplace**	**Garages**	**Pool**	**Lot size**	**Structure Size**	**Owner**
1	781700	49.18	3.792	4.374	0.2704	0.3377	0.1183	12540	2871	0.6007
2	266200	45.29	2.219	3.11	0.3488	0.5295	0.107	6737	1600	0.8928
3	633900	46.69	3.016	3.829	0.3546	0.452	0.1186	9773	2436	0.796
4	183200	48.71	6.108	6.884	0.2773	0.4195	0.134	24370	4917	0.4849
All obs	343600	45.68	2.438	3.305	0.3465	0.5092	0.1096	7591	1810	0.8633

Table 10: Mean Values for the Identified Submarkets.

	All observations		**Cluster 1**		**Cluster 2**		**Cluster 3**		**Cluster 4**	
Coefficients	**Estimate**	**Pr($>\|t\|$)**	**Estimate**	**Pr($>\|t\|$)**	**Estimate**	**Pr($>\|t\|$)**	**Estimate**	**Pr($>\|t\|$)**	**Estimate**	**Pr($>\|t\|$)**
(Intercept)	-3224	4.14E-11	10040	0.1091	-4319	$<$ 2e-16	-3096	1.8E-15	17050	0.1829
lot	3.68E-06	$<$ 2e-16	3.81E-06	2.75E-11	4.99E-06	$<$ 2e-16	4.68E-06	$<$ 2e-16	4.84E-06	$<$ 2e-16
structure	9.31E-05	$<$ 2e-16	1.73E-04	$<$ 2e-16	1.49E-04	$<$ 2e-16	1.48E-04	$<$ 2e-16	2.89E-05	5.97E-06
age	-0.0017	$<$ 2e-16	0.0008	0.3804	-0.0024	$<$ 2e-16	-0.0009	$<$ 2e-16	-0.0015	0.4875
bed	0.0093	4.68E-16	-0.0394	3.02E-08	-0.0033	0.0008	-0.0131	$<$ 2e-16	0.0343	1.78E-09
bathbed	0.1305	$<$ 2e-16	-0.0152	6.51E-01	0.0890	$<$ 2e-16	0.1122	$<$ 2e-16	0.1508	6.59E-09
garages	0.1150	$<$ 2e-16	0.1481	0.029872	0.2127	$<$ 2e-16	0.2532	$<$ 2e-16	0.2044	0.1293
pool	0.2418	$<$ 2e-16	0.3191	3.17E-07	0.1716	$<$ 2e-16	0.0988	$<$ 2e-16	-0.0529	0.6534
fireplace	0.2685	$<$ 2e-16	0.5422	2E-13	0.1874	$<$ 2e-16	0.2263	$<$ 2e-16	0.3891	0.0095
owner	0.0841	$<$ 2e-16	0.1777	0.0001	0.0980	$<$ 2e-16	0.0643	$<$ 2e-16	-0.2274	0.0220
Xcor	-24.7200	0.0021	240.4000	0.0209	-44.9000	$<$ 2e-16	-18.9200	0.0032	361.1000	0.0878
Xcor2	102.5000	$<$ 2e-16	1.0230	0.0201	-0.1880	$<$ 2e-16	-0.0772	0.0044	1.5280	0.0876
Ycor	-0.1020	0.0027	241.6000	1.55E-12	97.1300	$<$ 2e-16	114.9000	$<$ 2e-16	252.2000	0.0001
Ycor2	-1.5090	$<$ 2e-16	-3.5630	1.17E-12	-1.4290	$<$ 2e-16	-1.6930	$<$ 2e-16	-3.7070	0.0001
R2	0.3005	**no obs**	0.2665	**no obs**	0.5811	**no obs**	0.8127	**no obs**	0.3863	**no obs**
adjusted R2	0.3004	108488	0.2638	3636	0.581	85924	0.8126	17965	0.3779	963

Table 11: Regression Estimates

C Figures

C.1 IID case

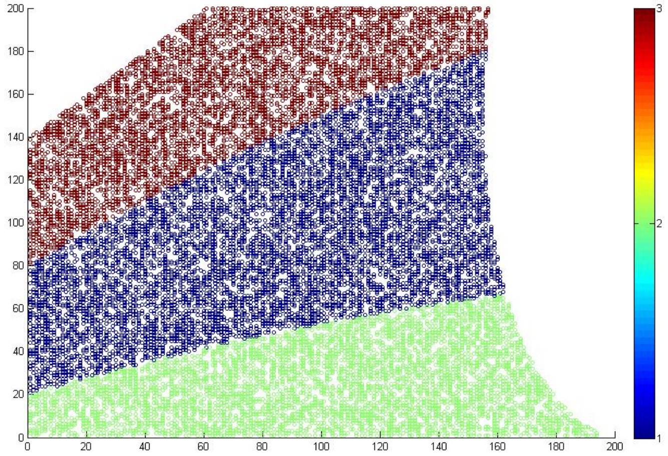

Figure 1: Low Variance

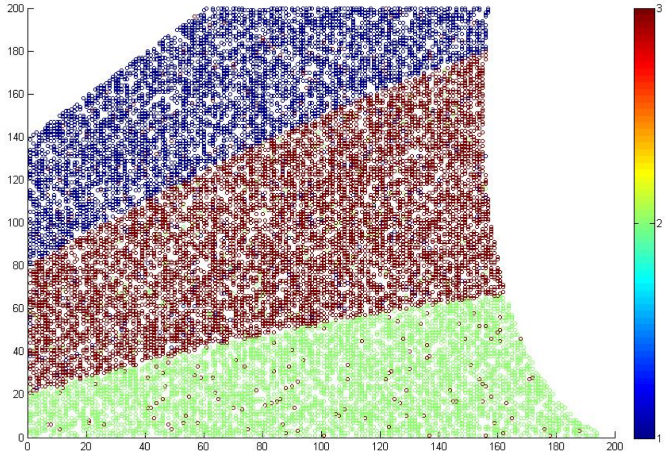

Figure 2: Medium Variance

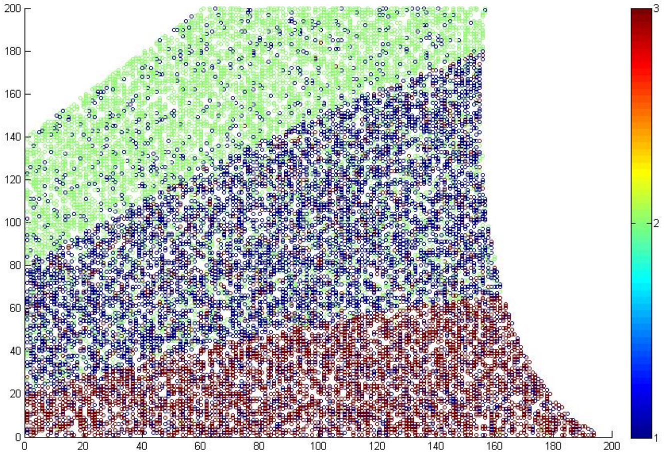

Figure 3: High Variance

C.2 Submarkets case

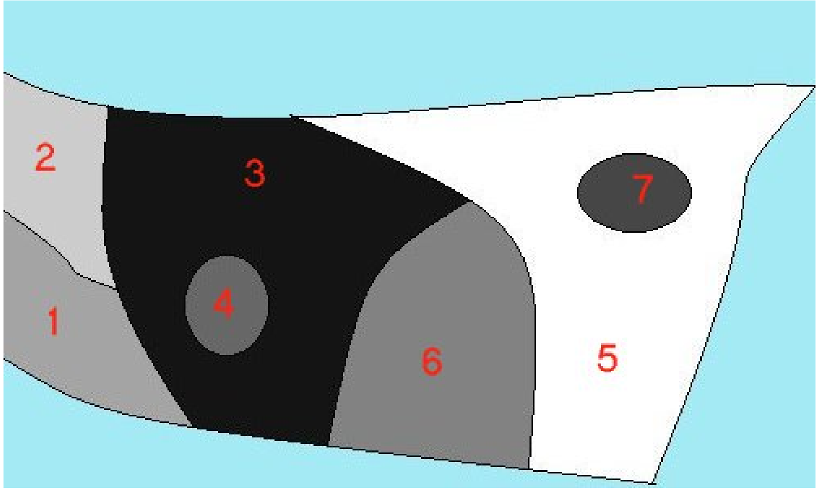

Figure 4: Complex Submarkets

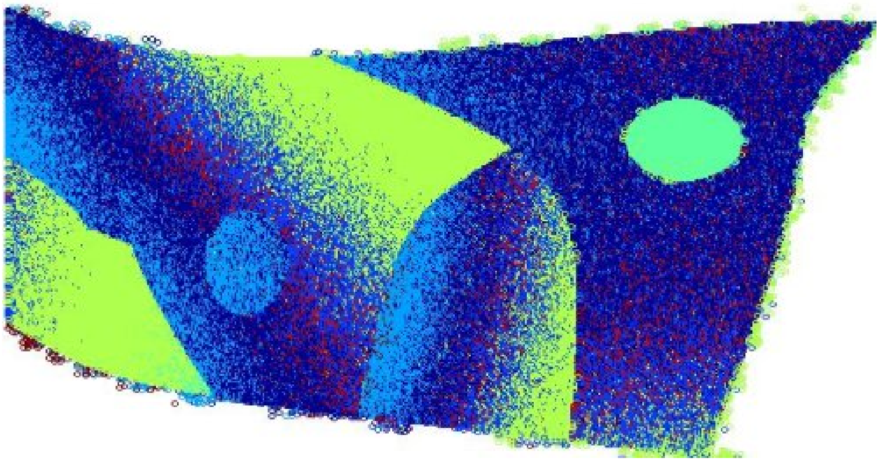

Figure 5: High Error Variance

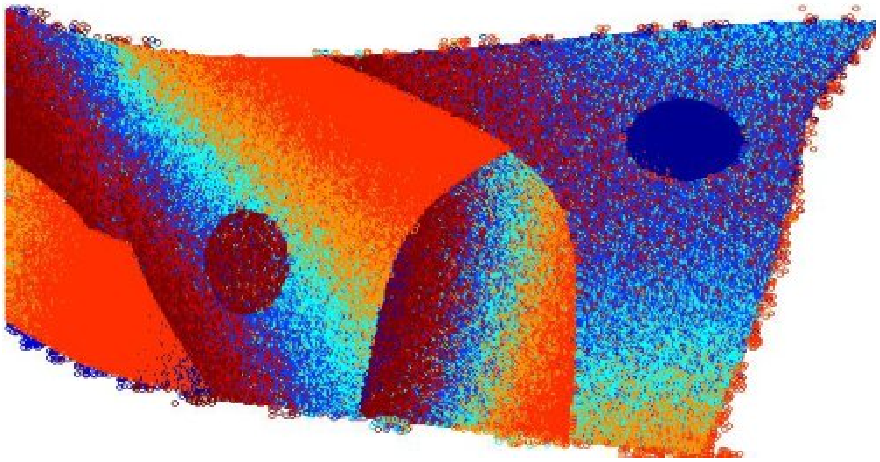

Figure 6: Low Error Variance

C.3 Real Data: Los Angeles

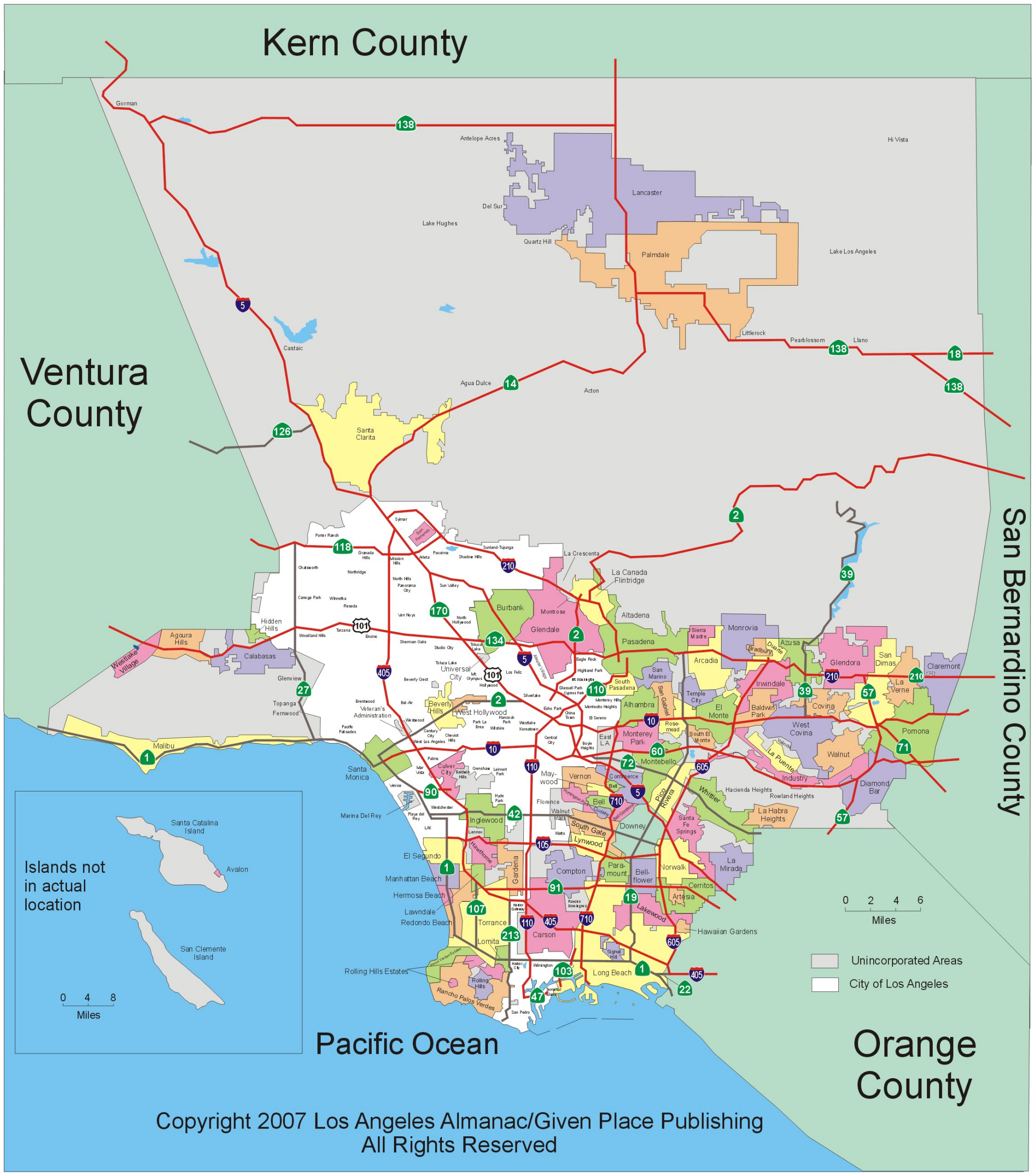

Figure 7: Los Angeles County Map

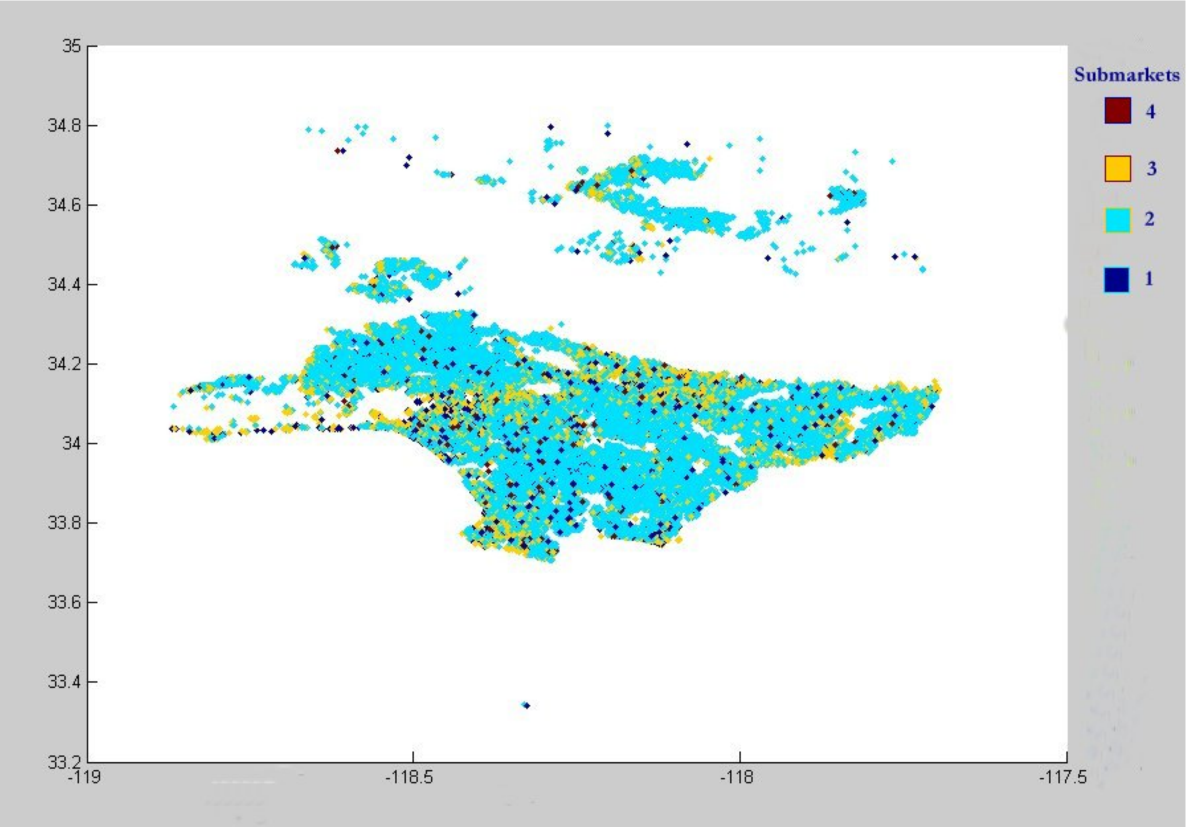

Figure 8: Los Angeles 2002 Submarkets

Research Papers
2010

CREATES
Center for Research in Econometric Analysis of Time Series

2010-39: Rasmus Tangsgaard Varneskov: The Role of Dynamic Specification in Forecasting Volatility in the Presence of Jumps and Noisy High-Frequency Data

2010-40: Antonis Papapantoleon and David Skovmand: Picard Approximation of Stochastic Differential Equations and Application to Libor Models

2010-41: Ole E. Barndorff–Nielsen, Fred Espen Benth and Almut E. D. Veraart: Modelling electricity forward markets by ambit fields

2010-42: Robinson Kruse and Philipp Sibbertsen: Long memory and changing persistence

2010-43: Dennis Kristensen: Semi-Nonparametric Estimation and Misspecification Testing of Diffusion Models

2010-44: Jeroen V.K. Rombouts and Lars Stentoft: Option Pricing with Asymmetric Heteroskedastic Normal Mixture Models

2010-45: Rasmus Tangsgaard Varneskov and Valeri Voev: The Role of Realized Ex-post Covariance Measures and Dynamic Model Choice on the Quality of Covariance Forecasts

2010-46: Christian Bach and Stig Vinther Møller: Habit-based Asset Pricing with Limited Participation Consumption

2010-47: Christian M. Dahl, Hans Christian Kongsted and Anders Sørensen: ICT and Productivity Growth in the 1990's: Panel Data Evidence on Europe

2010-48: Christian M. Dahl and Emma M. Iglesias: Asymptotic normality of the QMLE in the level-effect ARCH model

2010-49: Christian D. Dick, Maik Schmeling and Andreas Schrimpf: Macro Expectations, Aggregate Uncertainty, and Expected Term Premia

2010-50: Bent Jesper Christensen and Petra Posedel: The Risk-Return Tradeoff and Leverage Effect in a Stochastic Volatility-in-Mean Model

2010-51: Christos Ntantamis: A Duration Hidden Markov Model for the Identification of Regimes in Stock Market Returns

2010-52: Christos Ntantamis: Detecting Structural Breaks using Hidden Markov Models

2010-53: Christos Ntantamis: Detecting Housing Submarkets using Unsupervised Learning of Finite Mixture Models