Distinguishing Residential Submarkets

Gregory M. Schwann
Department of Property
University of Auckland

May 1998
Revised December 1998

Abstract

This article examines the possibility of dividing an urban housing market into a set of component submarkets in a way that is consistent with existence of one or more equilibrium hedonic pricing relationships. This is done by estimating an hedonic mixture model for the observed price. Two types of mixture model are estimated. The first is a standard fixed proportions mixture model, while the second is a spatially contextual mixture model. In both specifications, the null hypothesis of a single unified housing market is easily rejected. The exact number of submarkets is less clear, because some of the submarkets are too thin to obtain reliable regression estimates. Despite the results confirming multiple submarkets, my overall conclusion is that the prices for the vast majority of houses are described well by one or two equilibrium pricing relationships.

1.0 Introduction

This article is concerned with the possibility of dividing an urban housing market into a set of component submarkets. There is a growing literature on this topic. A central question in this literature is: What constitutes a (sub)market? In the literature, there are two basic approaches to defining housing submarkets. The most common approach is based on the notion that housing submarkets are composed of housing units which are close substitutes to each other, but which are poor substitutes for housing units in other submarkets (Grigsby, *et al.,* 1987; Rothenberg *et al.,* 1991). This definition of housing submarkets is expressed empirically through the classification housing units into submarkets based on their observed characteristics. For this reason, I refer to it as the market segmentation approach. Classification methodologies used to segment the market range from *ad hoc* definitions of submarkets to statistical methods, such as principal components, factor analysis or cluster analysis. The latter set of methodologies are based on a variety of housing and neighbourhood attributes, such as location, size, lot dimensions, view, etc. (Bourassa *et al.,* 1997; Dale-Johnson, 1987; Kendig, 1976; Kennedy, 1995; MacLennan and Tu, 1996; Smith and Kroll, 1989).

The other approach to defining submarkets is based on the premise that the law of one price must hold for all housing

units within a housing submarket; that is, units with the same attributes must have the same price. This means all housing units in a submarket have the same hedonic price ffinction. The corollary of this premise is that housing submarkets are distinguished by violations of the law of one price and, hence, differences in hedonic price ffinctions. This approach has not been used previously to define submarkets. Instead, it has been used to test whether a given segmentation of an urban housing market is valid. Examples of this procedure may be found in Bajic (1985), Dale-Johnson (1987), Gabriel (1984), Goodman (1981) and Schnare and Struyck (1976).

Both these approaches are incomplete. The market segmentation approach suffers from both theoretical and empirical problems. First, it does not take the product-differentiated nature of the housing market into account. The key point of the seminal article by Rosen (1974) is that, under suitable conditions, a competitive equilibrium can exist in a product-differentiated market, whereby products with different characteristics are in demand/supply equilibrium across a *unified* market. Note that substitutability of the differentiated products is subsumed within the hedonic pricing relationship generated by the competitive equilibrium. Hence, the substitutability of housing units only becomes an issue at the margin where the hedonic relationship breaks down. This complicates empirical work considerably, as market segmentation rules based on

housing attributes may not be linked in a meaningful way to housing submarkets. To resolve this problem, some form of market test needs to be built into the segmentation methodology.

The regression problem associated with the market segmentation approach is one of switching regressions with unknown, exogenous, sample separation. When the sub-sample obtained from the segmentation method contains observations from one submarket only, unbiased regression parameters can be estimated by ordinary least squares. This is the standard hedonic result. Empirical problems emerge **when housing units** are allocated incorrectly to submarkets. When misallocation occurs, the parameter estimates of the hedonic regression will be inconsistent (Aigner 1973, Lee and Porter 1984). The misallocation of observations also causes problems for hypothesis testing. The data in the segmented sample have a mixture distribution with unknown mixing proportions. Therefore, tests that rely on the (asymptotic) normality of the regression disturbances are invalid. This includes the usual Wald test for equality of parameters used to test whether a given segmentation of an urban housing market is valid. I do not know of any literature giving an approximating distribution for this test.

Obviously, existing applications of the law of one price approach suffer as much from the problems in hypothesis testing as does the segmentation approach.

In this article, I tackle the problems above by estimating a mixture model for housing transaction prices. The model endogenously distinguishes submarkets based on departures from an equilibrium hedonic pricing relationship(s). Thus, there is market test of segmentation. I estimate two forms of the mixture model. The first form is based on constant mixing proportions, while the second is based on the spatial assumption that houses close to each other are more likely to be in the same submarket than houses far apart. The spatial assumption of the second model is itself based on the theoretical proposition that households will sort themselves across space according to their housing preferences.

The paper is structured as follows. In section 2, I discuss the basic mixture model. In section 3, I examine the data used in my estimations. Section 4 contains a discussion of the results from my estimation of the mixture model. I develop the spatially contextual mixture model in section 5 and discuss the estimation results for this model. I conclude the paper in section 6.

2.0 The Mixture Model

I assume that there are M housing submarkets in the urban area. I examine the determination of the number of submarkets

in section 2.2. Each submarket is characterised by the market
conditional hedonic pricing model

$$p_{im} = x_{im}\beta_m + \varepsilon_m \qquad i=1,.--,I_m$$

(1)

where Pim measures the price of the ith housing unit in the
mth housing submarket, x_{im} is the $(1xk_m)$ vector of housing and
neighbourhood attributes for this housing unit, β_m is the $(k_m x\ 1)$
vector of regression parameters and ε_m is the regression
disturbance. The disturbances ε_m are independently and
identically normally distributed with mean 0 and variance $\sigma^2 m$
These distributions are defined only over the sub-population
of housing units in each market.

When the sample observations can be separated into submarkets
without error, the submarket pricing equations (1) can be
estimated efficiently by ordinary least squares. When sample
separation is unknown, bias, consistency and identification
problems emerge.

Suppose that one estimates the equations (1) by ordinary least
squares based on imperfect sample separation. Lee and Porter
(1984) show that the estimate β_m converges in probability to a
weighted average of the $\beta 1$, , β_M The weights in the average are
the conditional probabilities of classifying an observation
into submarket m given that the observation belongs in
submarket j. Consequently, the coefficients B_m, m=l,..,M, are
not estimates of the implicit prices of the dwelling
characteristics, but are estimates of an unknown mixture of
implicit prices. In addition to the inconsistency of the

parameters β_m, the estimates of the standard errors of the regressions, σ^2_m are biased and the direction of the bias can not be determined *a priori.* On its own, this confounds residual based hypothesis tests.

The inconsistency of the least squares estimate does not arise because of the definition of the submarket *per sea,* but from an inappropriate weighting of the observations in the subsample. Under least squares, an observation is given weight 1 when it is in the submarket sample and weight 0 when it is not in the submarket sample. The correct weight for an observation is proportional to the probability that the observation is in the mth submarket. Given the correct weights, weighted least squares yields consistent estimates of the parameters of the submarket hedonic regressions (1). Unfortunately, the weights must be estimated simultaneously with the regression parameters. When prior sample separation information exists, the parameters can be estimated using a switching regression model. However, with no prior sample

separation information (the usual case in hedonic estimation),
a finite mixture model must be used'.

In a finite mixture model, each observation on a house
transaction is viewed as a random selection from a population
of housing transactions generated by a mixture of M housing
submarkets. Therefore, the observed dependent variable, pj, is
the unconditional transaction price from the ith housing unit.
The marginal probability that an individual transaction is
drawn from the mth submarket is given by π_m where

$$\sum_{m=1}^{M} \pi_m = 1 \text{ and } \pi_j \geq j=1,. . .,M. \qquad (2)$$

These probabilities mirror the relative frequencies of houses
across the submarkets. The
probability density function for the ith observation is

$$\phi(P_i; \text{Xi}; \psi_m) = \sum_{m=1}^{M} \pi_m \phi_m \phi(>(P_i; \text{Xi}; \psi_m) \qquad (3)$$

where $\phi((P_i; \text{Xi}; \psi_m)$ is the normal probability density function
for the ith observation, when it is drawn from the mth housing
submarket. In equation (3), the vector xi contains the housing
characteristics used in all the housing submarket regressions
(1). In many applications, the same housing attributes will be
used in each submarket regression and xi=xim for all m. In
other applications, the attributes in the hedonic regressions

(1) may vary across the submarkets and x; will be the union of submarket specific attributes. For example, a particular geographical grouping of housing units may need extra view or location attributes to describe the properties. Alternately, there may be no variation in one or more attributes for houses in a submarket and, hence, the zero variation attributes must be deleted from the attribute set. The effect of the zero variation attributes is impounded in the constant term. Also in equation (3) $\psi_m = (\beta_m \sigma^2_m)$ is the vector of parameters for the mth submarket and $\psi = (\pi_1, \ldots . \pi_M; \psi_1 - .. \psi_M)$ iS the full parameter vector for the mixture distribution.2

Since the observations on the transaction prices of housing units are independent and identically distributed, the log-likelihood, L(v), is the sum of the logarithms of the probability density functions for the observations in the sample. An estimate of v can be obtained as the solution to the equations

$$\partial L(\psi)/\partial \psi = o . \qquad (4)$$

For the mixture model above, these can be analytically determined (see Lee and Porter 1984). As already noted, the estimates of the regression parameters ~m are weighted least squaresestimates. Similarly, the estimate of the submarket variance smis the weighted sum of

squares divided by the expected degrees of freedom for the regression. For both sets of parameters, the weights are the

posterior probabilities of submarket membership derived from the mixture distribution. For a given estimate of the parameters ~, these posterior probabilities are given by

$$\iota_{im} = \pi_m \phi_m(P_i; X_{im}; \Psi) \; / \; \sum_{m=1}^{M} \pi_j \phi_j(P_i; X_{im}; \Psi) \tag{5}$$

where $\iota_{im} = pr\{observation\ i\ belongs\ to\ submarket\ m | x_i, \Psi\}$. These estimated posterior probabilities may be used to segment the sample into submarkets according to the modal rule $m_i = argmax\{\iota_{im} l m = 1, ..., M\}$.

Last, solving $\partial L(\psi) / \partial \pi_m = 0$ gives the estimates

$$.... \qquad \pi_{m\ =} \sum_{i}^{N} \iota_{im} \; / N \qquad m = 1, ...M \tag{6}$$

for the prior probabilities of submarket membership.

Over time, a number of authors have noted the remarkably simple structures of the estimators. These authors have suggested a simple iterative solution for maximising the log likelihood and obtaining the parameter estimates. All of the suggestions are variants of the EM algorithm of Dempster, Laird and Rubin (1977). A complete discussion of the EM

algorithm and its application to mixture models is contained in McLachlan and Basford (1988) or Titterington, Smith and Makov (1985). The most compelling feature of the EM algorithm, beyond its simplicity, is that log-likelihood value is monotonically increasing. Unfortunately, this is balanced by very slow convergence, particularly in the region of the root of likelihood. I use the EM algorithm to obtain initial estimates of the parameters, but refine the solution using Newton-Raphson.

2.1 Parameter Identification

In formulation above, all of the densities 4)m are normal. This creates an initial identification problem, because the mixture density _ will have the same value when the submarket labels m and n are interchanged in equation (3). In fact, since there are M! permutations of the submarket labels, there are M! models which yield the same value of _. This problem can be
resolved quickly by imposing identification restrictions that order the submarkets. Aitkin and Rubin (1985) suggest imposing the restrictions zl27r22.. .2xM>O. In practise, this or any similar ordering restriction does not need to be imposed while estimating the parameters. The submarket indexes can be ordered as required after solving for the parameter estimates. A more pressing identification problem was first noted by

Kiefer and Wolfowitz (1956) in the context of a two component mixture. They showed that the log-likelihood for a mixture of univariate normals with unequal variances is unbounded.3 Hence, a global maximum likelihood estimate does not exist. Kiefer (1978) and Hathaway (1985) extended this result to more general mixtures. The lack of a global maximum does not stop one estimating the mixture model. There exists a sequence of roots to the likelihood equation (4) that is consistent and asymptotically efficient under suitable regularity conditions. In fact, the mixture model may have multiple roots, with each root corresponding to local maximum of the log likelihood. Because of this, one should estimate the mixture model from a variety of starting points.

Alternately, identification restrictions can ensure the existence of a well identified solution. To derive the necessary restrictions, first note that the log-likelihood has a global maximum when the variances of the components are equal. Heuristically, it appears that restrictions on the variances of the regressions are needed to bound L. Hathaway (1985) demonstrates that a constrained global maximum exists when the parameter space is restricted to the region defined by

6,2a622...2a6M ,, and 6M2a61 for any 0<aS1 Observe that the constraints are not very restrictive; they just ensure that the variances are finite for every submarket. If the true value of the mixture parameters v also lies in the constraint

set, then the constrained global maximum is a strongly consistent estimator for v and is the largest local maximiser of the unconstrained likelihood maximisation problem.

I do not impose Hathaway's restrictions in my estimations. Nevertheless, the estimated variances from my models satisfy the restriction when $0 < a \leq .17$. Since a=1 corresponds to an equal variance model, an upper bound of a=0.17 demonstrates the necessity of expanding the range of the parameter space when dealing with housing submarkets.

2.2 Tests for the Number of Submarkets

There is no definitive test for the smallest number of submarkets, M, in a mixture of markets.

The likelihood ratio test statistic X is obvious candidate for such a test. Unfortunately, the standard regularity conditions required for -2 lnk to have an asymptotic chi-squared distribution with degrees of freedom equal to the number of restrictions in the test are violated in a mixture model. To understand the basis of these violations, consider testing the null hypothesis of M versus M+1 submarkets. This can be accomplished by imposing a single restriction, 7tM+I=0 or by imposing the kM+g+1 restrictions pM+I=0 and 62M+l=0. The

first approach places 1tM+l on the boundary of the parameter space. This is a direct violation of the regularity conditions. Under the alternate set of restrictions,7rM+l iS irrelevant, but even here the usual regularity conditions do not hold, but for more subtle reasons (Ghosh and Sen, 1985; Quinn and McLachlan, 1986).

Despite the theoretical problems with the likelihood ratio test, the test is still used as a guide to the possible number of submarkets in the mixture (McLachlan and Basford, 1988). The use of the test is supported by several Monte Carlo studies. Wolfe (1971) examines the test of the null hypothesis of M=1 versus M=2 groups in a mixture of normals. He shows that the null distribution of 2c InS is approximately chi-squared with degrees of freedom equal to twice the number of restrictions (not including the mixing proportions) in the test. The constant c=(n-l/2m-2)/n, which is approximately equal to unity for large n.. Hernandez-Avila (1979) and Everitt (1981) confirm this result.

Alternately, one might test the null hypothesis using the bootstrap distribution of -2 ln B. Efron and Tibshirani (1986) show that a large number of replications are needed to obtain accurate estimates of P-values. Even for simpler models than mixtures, over 350 replications are needed. While I agree with the advisability of calculating the bootstrap distribution, I do not do so because of the computational burden.

3.0 Data

The mixture model is estimated using transaction data for the City of Auckland, New Zealand. These data are drawn from the Valuation New Zealand (VNZ) transactions database and marketed by Headway Systems Ltd. VNZ is the federal government agency charged with maintaining the Torrens land registry for New Zealand. All property sales are registered with VNZ. The Headway Systems database does not contain all the fields in the VNZ database. What are missing are variables describing the interior layout of the dwelling, for example, the number of bedrooms and the number of bathrooms.

The data used here are for detached residential units sold between 1995:Q4 and 1996:Q3. This was a time of high sales volume and it marked the end of a period rapid price appreciation. There are 4,712 observations in the sample.

I present the summary statistics for my data set in Table 1. The average price of the housing units sold in 1996 was $307,592. The range of transaction prices underlying this mean price is extremely wide. The least expensive housing unit sold for $2,500; the most expensive housing unit sold for $3,250,000; and the standard deviation was $190,330. I explain this house price variation using the nine housing attributes in Table 1. The rest of this section is devoted to a presentation of the definitions of these housing variables and a discussion of possible limitations of my measures.

The variable BUILT measures the date of construction, but not perfectly. The data file contains only a set of indicators for the decade of construction. The value of BUILT is the first year in the decade. BUILT is a proxy for pure depreciation and, in this role, it should have a positive effect on price. However, the BUILT variable is also a proxy for two other effects that confound its impact on price. First, because vintage of a housing unit is linked to its architectural style, BUILT captures the value of current household preferences for architectural style. This value of vintage related architectural styles could be either positive or negative. Anecdotal evidence suggests that some of the older styles are currently in demand and this will impart a negative effect to the BUILT variable. Second, like many new cities, Auckland was developed in successive waves emanating from its historic urban core. As a result, BUILT also measures distance from the CBD and this effect also imparts a negative sign. However, Auckland is now a multi-nucleated urban region, so the importance of distance to the CBD effect is not clear.

Table 1

Descriptive Statistics

Variable	Mean	Std	Minimum	Maximum
PRICE	307591.72	190330.65	2500.00	3249999.57
BUILT	1945.78	25.86	1880.00	1990.00
FLR	135.14	66.00	10.00	1990.00
LAND0	0.31	0.46	0.00	1.00
LAND	778.34	428.59	0.00	2431.00
CHATTELS	13556.95	8522.51	0.00	89000.02
AVER	0.79	0.41	0.00	1.00
POOR	0.06	0.24	0.00	1.00
T	0.35	0.21	0.00	0.88

FLR is the floor area of the housing unit in square meters. The average house sold in Auckland in 1996 was 135 square meters (approximately 1,450 square feet). This is considerably smaller than the new homes constructed in the Auckland region, which are about 210 squares meters (approximately 2,300 square feet).

LAND is the area of the lot in square meters. The average lost size was 778 m2 (approximately 8,400 square feet). Unfortunately, 31% of the transactions in the data file does not contain the lot area.4 Moreover, there is no a clear reason for the missing data. Rather than eliminate these transactions, I include the variable LAND0 in the regressions. LAND0 is an indicator variable which is assigned the value unity when the lot area is not recorded and is assigned the value zero when the lot area is recorded. In using this specification, I am assuming that the marginal prices of the

other attributes do not depend on whether the lot area was recorded.

The variable CHA'l l'ELS is the value of chattels in the dwelling unit. Since CHATTELS are mobile capital, the value of chattels should be fully capitalised into the price of the housing unit. On this basis, the house price elasticity of CHA'l'l'E;LS should equal their relative share in value, or approximately 4.4% in my sample. In past work with this data set, I have obtained much larger positive elasticities. This result suggests that CHATTELS is a proxy for the general quality of the interior space of the dwelling.

AVER and POOR are indicator variables which measure whether the unit is in average or poor condition. The omitted category is good houses. Thus, both AVER and POOR should exert a negative effect on house price. These variables are included in the Valuation New Zealand database. They are based on the professional assessment of the Valuation New Zealand valuers (*sic* appraisers).

The final variable is a time index, denoted by T. It equals the number of days from the October 1, 1995. It is included to capture inter-year price appreciation. The variable T is included because there was considerable price appreciation in the Auckland housing market in the time period of under examination. In the estimations, all the variables are log transformed and a log-log model is estimated.

4.0 **Mixture Model Results**

The mixture model described in section 2 separates the urban housing market into submarkets using the residuals from the hedonic models (1). The difficulty of this problem is highlighted in Figure 1. The map in Figure 1 shows the residuals from the single market hedonic model of the housing transactions. The residuals are divided into three groups: strongly negative residuals (large dots), which are more that 1 standard error below zero, "neutral" residuals (small dots), which are between -1 and 1 standard errors from zero, and strongly positive residuals (large triangles), which are more that 1 standard error above zero. The figure shows that the vast majority of housing transactions are fit well by a simple model ($R^2=0.81$). This supports the view that housing markets are unified product-differentiated markets. The figure also shows that there is no clear spatial pattern to the strongly negative or strongly positive

residuals. No one neighbourhood has consistently negative or positive residuals, although there are sub-neighbourhood clusters of negative or positive residuals. This ocular review indicates that there is no simple model for defining submarkets.

I estimate the mixture model for an urban housing market containing from 1 to 5 submarkets. I present the log likelihood trace for this set of estimation experiments in the first column of the left-hand panel of Table 2. The likelihood ratio test statistic -2 In x for the tests of the

null hypothesis of m versus m+l housing submarkets are presented in the second column of the panel.I can reject the null hypotheses of 1 submarket versus 2 submarkets, the null hyppthesis of 2 submarkets versus 3 submarkets and the null hypothesis of 3 submarkets versus 4 submarkets using either a conventional chi-squared distribution for the test statistic or Wolfe's adjusted chi-squared distribution with twice the number of degrees of freedom (Wolfe's c~1). My attempts to estimate a 5 submarket model failed. Although the EM algorithm does produce estimates for the 5 submarket model, different estimates are obtained from different starting values and covariance matrix of the parameters for these estimates is not positive definite. A deeper investigation of the fifth market segment reveals that it has too few observations to permit a reliable estimate of the hedonic model for that submarket. A model with a reduced set of explanatory variables might be estimated, but I have not taken

that step here.

Table 3 contains the estimated coefficient for the 4 submarket model. The means for the variables for the observations in each submarket are given in Table 4, to assist in the interpretation of these estimates. To calculate the means, the observations are assigned to the

submarket with to the largest posterior probability; that is,

$m_i = \arg \max m \{\iota_{im}\}$

Table 2

Log Likelihood Trace For Mixture Models

Submarket	Oure Mixture		Spatially contextual mix	
	LogLiklihood	-2lnλ	LogLiklihood	-2lnλ
1	-1930.98	--	-1930.98	--
2	4428.91	12719.79	4437.98	12737.93
3	4553.29	248.785	4532.97	189.97
4	4792.81	479.05	4591.49	117.05
5	nc	nc	4637.61	92.24

nc = Model failed to converge.
Critical value for chi-squared with 10 degrees of freedom is 18.3.
Critical value for Wolfe's chi-squared with 20 degrees of freedom is 31.4.
(df=lO)

The most significant feature of the estimates is that prior probabilities of submarket membership, x, allocate the majority of the observations to one or two submarkets. Sixty-six percent of the observations are in the first submarket and twenty-three percent of the observations are in the second submarket. The corresponding posterior

probabilities are 72.1% and 14.9%. While the hypothesis of a single housing market is rejected, quite resoundingly, the fact that 89% of the observations lie in one or two submarkets supports the view that relatively few product-differentiated markets are needed to describe pricing in urban housing markets.

The parameter estimates for the hedonic equations vary quite widely across the submarkets. This is expected. If they did not differ significantly, the model would not discern different pricing relationships.

The coefficients for submarket 1 point to a market with two dominant pricing factors: the vintage of construction and the value of chattels. The elasticity of the vintage variable, BUILT, is -0.232. The fact that the coefficient is negative means that in the Auckland market the vintage of the housing unit and/or its proximity to the CBD outweigh the pure depreciation effect of unit ageing. However, despite the magnitude of the elasticity, the actual effect in the market is tiny. At the margin, a new 1995 vintage house will be only 0.5% cheaper than the average 1946 vintage house. The other noted pricing factor, CHATTELS, does have a demonstrable effect on price. The elasticity of price with respect to CHATTELS is 0.946, which translates to a 6.5% increase in property value per $1,000 of chattels (evaluated at the mean for submarket 1). Thus, the general quality of the interior

space of the dwelling is the overriding factor explaining the prices of the housing units in this submarket. Although the other pricing attributes are statistically significant (except for POOR and T), the estimated elasticities are small and the quantitative impacts negligible. As expected, house prices increase with floor area and decrease when quality declines, but, unexpectedly, house prices decrease with the lot size variables LAND and LAND0. The negative sign on the LAND coefficient implies that, at the margin, the average lot in this submarket is too big. The negative sign on the LAND0 reflects the negative coefficient on LAND. An average property would have a LAND0 coefficient of about -0.53. The estimated coefficient for LAND0 is -0.55, which is not significantly different from -0.53.

=== Tables 3 & 4 ====

The housing units in submarket 2 (see Table 4) are cheaper- $258,125 versus $308,602, with less value in chattels - $13,026 versus $14,119, but on larger lots _ 793m2 versus 762m2. Also, the number of properties with no recorded lot size increases from 30% to 44%. The estimated price elasticities show that, like submarket 1, the value of chattels is the most important attribute in housing unit pricing. The elasticity of CHATTELS is 1.083. This factor alone accounts for $21,562 of the $50,477 difference in the mean prices of the units in the two submarkets. The pricing of units in submarket 2 differs from the units in submarket 1 in

three respects. First, in submarket 2, there is a greater reliance on lot size as a determining factor. This is manifest primarily through an increase in the LAND coefficient. A 100 m2 increase in lot size increases the price of the property by 4.2%. Second, housing unit quality is more important in submarket 2. Average units are 5.6% less expensive than good units, as compared to 1.1% less expensive in submarket 1. The coefficient for the variable POOR is statistically insignificant in both submarkets. Third, the vintage of the housing unit has a smaller positive effect on price in submarket 2 and it is statistically insignificant. Submarkets 3 and 4 are minor markets, constituting 6.2% and 4.8% of the Auckland City housing units, respectively. Both are higher priced markets with average sale prices of

$379,170 and $320,625 relative to the $308,602 and $258,125 for submarkets 1 and 2. The higher prices are supported by dwelling units that have more floor area, that are located on larger lots and that are of higher quality, but the value of chattels is lower and this partially offsets the impact of the favourable attributes.The offset from lower chattel values is smaller than it would be in the first two submarkets, because the housing prices in submarkets 3 and 4 do not depend as heavily on this pricing factor. The elasticity for CHATTELS is 0.029 for submarket 3 and 0.177 for submarket 4..

Submarket 3 is sensitive to the vintage of construction. The

prices of housing units in this submarket decrease by 4.3% per decade. The effect of BUILT is statistically insignificant in submarket 4. In both submarkets 3 and 4, the prices of the units increase markedly with floor area. The elasticities of FLR are 0.813 and 0.507, respectively. These coefficients imply that an increase of 10m2 (approximately 108 square feet) increases the house prices in these submarkets by 5.28% and 3.23%, respectively. Lot size does not have a statistically significant effect on the prices in either submarket. Finally, the housing units in these submarkets respond very strongly to decreases in housing quality. An average quality unit in submarket 3 is 45.6% less expensive than a good unit and in submarket 4 an average unit is 36.3% less than a good unit. A further decrease in quality to poor from average reduces the price in submarket 3 by an extra 0.23% for a total decrease of 47.9%. The corresponding decrease in price submarket 4 is 35% for a total decrease of 71.3%, but this marginal effect is not statistically significant.

5.0 A Spatially Contextual Mixture Model

One limitation of the mixture model presented in section 2 is that the mixing probabilities Z are constant. As a result, the ability to distinguish submarkets depends solely on the separation of the distributions of the regression disturbances eim. This approach does not use all of the market information available. In particular, it does not use the fact that household will sort themselves across space according to their housing preferences.5 A well-known example is the sorting of household by size and/or composition. Households with children demand more land relative to housing capital. Hence, they have flatter bid rent curves and, in a competitive housing market, this results in their locating further from employment nuclei (further from the CBD in monocentric cities), other factors being equal. Another well-known example is that of households sorting themselves, *a' la* Tiebout, by jurisdiction according to their personal net fiscal benefit of residing in a community.

The gist of household sorting is that like households will congregate in neighbourhoods. It is the essential idea of congregation or neighbourhoods that must be captured in any empirical specification.6 In the rest of this section, I specify a reduced form model that captures the idea of spatially contextual submarkets and present the estimates from this model. The model is based on the hypothesis that housings unit which are spatially close to each other are more likely

to be in the same submarket than housing units that are spatially distant from each other. This hypothesis already has some empirical support in terms of the subneighbourhood clusters of negative and positive residuals observed in Figure 1. A formal test is needed however.

To define the contextual model, I need to develop some notation. Let $z_i = (0,\ldots,0,1,0,\ldots,0)$ be an M dimensional submarket indicator vector for the ith housing unit, where $z_{im} = 1$ if housing unit i is in submarket m and is zero otherwise. Also, let $z = (z_1,\ldots,z_1)$ be the collection of the indicator variables for all the housing units in the urban housing market. The problem of determining a spatially contextual estimate of submarket membership is one of specifying the distribution for the $pr\{z = z \mid p, x\}$. This distribution is the joint distribution of submarket assignments across all housing units conditional on the pricing information in the market.

To estimate z, the probabilities of submarket membership must be estimated simultaneously for all housing units. This is a daunting computational task even for small markets. Fortunately, methodologies for doing this have been explored in the literature on statistical image analysis. In that literature, the problem is to assign colours to the pixels in an image contaminated with noise. This problem occurs in the remotely sensed images from satellites, medical imaging (e.g., magnetic resonance or photon tomography) and astronomy, to

name a few examples.

Models based on locally dependent Markov random fields play a central role in statistical image analysis and correspond closely to the problem of assigning housing units to submarkets . When the probability distribution $P(z)=$ prt z = z I p, x } follows a Markov random field, each housing unit can be assigned to a submarket, according to the conditional probability

Pi (Zi I Z(i)) = Pr{Zi = Zi I Z(i, } …..(7)
where Z(i) is the collection of submarket allocations z with zi deleted. The additional assumption of local dependence means that the distribution (7) relys only on the housing units in the neighbourhood of housing unit i. Let Ni be a subset of { 1,...,I} containing the indexes of the neighbours of housing unit i. Then, the conditional probability (7) becomes

Pi (Zi | ZNj) Pr{Zi Zi | ZNj } (8)

To use this approach, a functional form and neighbourhood structure must be specified a *priori*. Besag (1986) and others suggest the simple, frequency based, functional form

prt Zim = 1 1 ZN, } = Zm exp(]/u im)/~jMI 1tj exp(^yu jj) m = 1,. . ., M; i = 1,.. ., I (9)

where uim denotes the number of neighbours on housing unit i belonging to submarket m.

The coefficient Y measures the strength of the submarket relationship among the housing units in the neighbourhoods, Ns. When >0, there is no relationship between the submarket assignments and the model reduces to the pure mixture model. As y increases the correlation between the submarket assignments increases.

Defining the neighbourhood of a housing unit is a more difficult task than it appears at first. Most choices for a neighbourhood do not yield a closed form solution of (9). The difficulty lies in the fact the neighbourhoods of dwelling units overlap and, therefore, one still must solve (9) simultaneously for all housing units. The exception to this are Markov mesh models (Abend, *et al.,* 1965), which impose a pre-specified order of precedence on the evaluation of the probabilities. In the empirical work below, I use a Markov mesh model in which the neighbourhood of housing unit i contains all housing units within a one-third kilometre radius of unit i and lying to the south-west of this unit. Clearly, this is a restrictive assumption but it is needed to simplify the computation.

The spatially contextual mixture model is obtained by substituting the probabilities (9) for the fixed mixing proportions in equations (3), (5) and (6). The parameter estimates are solution to (4). A version of the EM algorithm can be used to estimate the model. However, for convenience, I estimate the model using a combination of the BHHH algorithm

and the Newton-Raphson algorithm. I present my results for this model in table 2, 5 and 6.

The log likelihood trace for the spatially contextual model is recorded in the right hand panel of Table 2. The log likelihood ratio tests support a model containing five housing submarkets. An attempt to fit a model with six housing submarkets failed. The model allocated too few observations to the sixth market to enable the estimation of the parameters of the hedonic equation for this market.

There is one anomalous result in Table 2. The log likelihood for the contextual mixture model is less than the log likelihood for the pure mixture model for each submarket dimension, except m=2. Normally, adding an extra parameter increases the log likelihood value. The decrease in the log likelihood is due to either the neighbourhood definition or the Markov Mesh assumption. In the future, additional specifications should be examined. The estimated coefficients for the model are given in Table 5 and the means of the variables for the five submarkets are presented in Table 6. The first coefficient of interest in the model is r, since it defines the strength of neighbourhood relationships. My estimate is y=0.065.

The estimate is statistically significant, but numerically small. However, despite its size, it has a large effect on the estimates of the other coefficients in the model. By introducing a spatial context, housing units are bound

together into neighbourhoods and neighbourhoods are bound together into an extensive urban housing market. This is manifest in the model by the high proportion of housing units allocated to the dominant housing market_submarket 1. The prior probability that a house will be assigned to submarket 1 is 83.5% and the posterior probability is 89.5%. The corresponding values for the pure mixture model where 65.7% and 72.1%.

=== Tables 5 & 6 =

The estimates of regression parameters for submarket 1 are qualitatively similar to the estimates for submarket 1 in the pure mixture model. Once again, CHATTELS is the most important price determinant in the dominant submarket, with an elasticity of 0.972. The variable BUILT, which was important in the pure mixture model, is smaller in magnitude and statistically significant in the contextual model. In fact, BUILT is only statistically significant in two submarkets: submarkets 2 and 5. This suggests that adjusting for neighbourhood context reduces the effects of architectural vintage and distance from the CBD. The floor area variable, FLR, has a small positive and statistically significant coefficient, as it does in the pure mixture model. The lot area variables LAND0 and LAND have small negative coefficients. However, the estimates are statistically insignificant. The insignificance of the lot area variables is

a general result. The coefficients are insignificant for all submarkets, except submarket 5. In the pure mixture model, LAND0 and LAND are statistically significant for the two dominant markets. The elasticities of the quality variables AVER and POOR are larger in magnitude and statistical significance for all of the submarkets than in the pure mixture model. In submarket 1, an average units is 1.2% cheaper than a good unit, while a POOR unit is 1.9% cheaper than a good unit. The quality effects are more pronounced in the other markets.

Submarket 2 corresponds roughly with submarket 3 the in the pure mixture model. In both models, it is consists of big expensive houses. There are two main differences between the models. First, the elasticity of FLR is smaller in the contextual model than in the mixture model (0.737 versus 0.813). Second, the quality variables are also slightly smaller in magnitude than contextual model than in the mixture model (-0.407 versus -0.456 for AVER and -0.446 versus -0.479 for POOR). The pricing effects of the decreases in the elasticities is muted by the increase in the average floor area and increase in general quality of the houses in submarket 2 of the contextual model as compared to submarket 3 of the pure mixture model.

Submarkets 3, 4 and 5 have no clear analogues in the pure mixture model. All three markets contain expensive houses, with average sales prices of $384,106, $315,467 and $530,913,

respectively. Moreover, all three submarkets display remarkable estimated rates of price appreciation. Over the time period starting in 1995:Q3 and ending in 1996:Q3, these housing units appreciated by 42.6%, 47.7% and 57.5%. These rates of appreciation are statistically significant for submarkets 3 and 5, but not for submarket 4.

Submarkets 3 and 4 are similar in terms of their housing attributes and in terms of their estimated elasticities. Submarket 5 is quite distinct. The houses in submarkets 3 and 4 are about the same average age, built in 1946 and 1945; they are large on average, 176 m2 and 171 m2and they are located on large lots, 877 m2 and 855 m2 on average. They differ in the average value of their chattels, $12,988 and $9,362, and the average quality of the units, The coefficients on the BUILT variable are large but statistically insignificant in both markets. The LAND0 and LAND variables are also statistically insignificant in both markets. Although, I note that estimated coefficients for the lot size variable are strongly negative for submarket 3. The prices of the dwellings in these markets respond strongly, and with statistical significance, to the floor area of the houses. For submarket 3, the elasticity of FLR is 0.906, while, for submarket 4, the elasticity of FLR is 1.431. The elasticities of CHATTELS for the submarkets are 0.144 and 0.254, respectively, and both coefficients are statistically significant. These elasticities are smaller than those observed for the markets in the mixture model. The

quality variables have major impact on the prices of housing units in submarkets 3 and 4, particularly the POOR variable. In submarket 3, housing units are marked down by 19.2% if they are average quality units and 66.3% if they are poor quality units. Both coefficients are statistically significant. In submarket 4, the percentage reductions in price are 6.3% and 94.0% for average and poor quality units respectively. However, only the coefficient on POOR is statistically significant. It is hard to discern precisely what differentiates the pricing of housing units submarkets 3 from the pricing of units in submarket 4. But, in general, it appears that the prices of the units in submarket 4 are more attribute responsive than the prices of the units in submarket 3.

Submarket 5 is unique among the submarkets. It contains a small number of very expensive newer, smaller, houses built on large lots. In addition, the houses contain chattels of considerable value. The properties are priced quite differently from those in the other submarkets. The estimated elasticity on BUILT is large_10.271_.and, as noted above, this is the only submarket where BUILT is statistically significant. The estimated elasticity implies that the houses in this market rise by 5.2% for every decade newer. Despite the smaller sizes of these housing units, the elasticity of price with respect to floor area is small and statistically insignificant. On the other hand, despite large lot sizes, the

prices of the houses are quite sensitive to lot size. The elasticity for lot size is 1.099, which implies a price increase of 9.6% per 100 m2. This strong lot size effect is reflected in the LAND0 coefficient. If the lots without a recorded land area were average in lot area, the LAND0 coefficient would be approximately 7.6. The estimated coefficient is 7.38. The opposing results for FLR and LAND variables suggest that the properties in this submarket are prime for subdivision and redevelopment. The large positive and statistically significant coefficient on the quality variable POOR (0.955) is further evidence supporting this conjecture. It is easier to obtain a the necessary planning approvals when the property is run down. The positive elasticities on BUILT and CHATTELS and the negative elasticity on AVER might be considered counter evidence. These coefficients indicate that housing unit attributes have a considerable role in determining price. However, this argument does not have much force in Auckland, because of the ease and prevalence of in-fill redevelopment, wherein the existing structure is retained and the "surplus" land is subdivided and built upon.

6.0 Summary and Conclusion

This paper demonstrates that it is possible to divide an urban

housing market into a set of component submarkets in a way that is consistent with existence of one or more equilibrium hedonic pricing relationships. The test for the number of component submarkets easily reject the null hypothesis of a single unified housing market, given the hedonic models specified in the paper. The exact number of submarkets is less clear. The models run into thin submarket problems before the log likelihood ratio tests can reject the null hypothesis of an additional submarket. Despite the results confirming multiple submarkets, my overall conclusion that the prices for the vast majority of houses are described well by one or two equilibrium pricing relationships. The remaining submarkets are niche markets of some sort. Unfortunately, the houses that fall into these niche markets are not identified in advance.

My results suggest that identifying the houses that fall into in the smaller submarkets is not the appropriate objective. Instead, if the objective is to fit the equilibrium hedonic relationship for the dominant submarket, this can be accomplished by an aggressive program of data trimming. If the houses in the smaller submarkets are removed, the hedonic price function for the dominant submarket can be consistently estimated by ordinary least squares. For Auckland City, trimming by price, floor area and lot area would yield the requisite sample.

Table 3

Parameter Estimates by Submarket for Mixture Model

Variable	Submarket							
	1		2		3		4	
	Coeff(t)	t statisti	Coeff(t)	t statisti	Coeff(t)	t statisti	Coeff(t)	t statisti
P	0.657	33.746	0.233	10.521	0.062	7.971	0.048	6.034
Constant	5.297	11.301	1.990	1.035	72.285	4.674	-16.763	-0.368
BUILT	-0.232	-3.982	0.014	0.053	-8.384	-4.145	3.183	0.521
FLR	0.020	5.131	0.006	0.737	0.813	7.840	0.507	2.425
LANDO	-0.055	-2.364	0.216	2.105	-0.080	-0.134	1.292	0.828
LAND	-0.008	-2.383	0.035	2.227	-0.003	-0.035	0.222	0.951
CHATTELS	0.946	230.308	1.083	81.548	0.029	3.508	0.177	10.083
AVER	-0.007	-2.296	-0.039	-3.445	-0.456	-4.998	-0.363	-2.019
POOR	-0.011	-1.872	-0.056	-4.113	-0.479	-3.225	-0.713	-1.786
T	-0.005	-1.037	0.024	1.579	0.187	1.542	0.440	1.585
S	0.043	41.663	0.057	26.668	0.252	5.551	0.752	11.069

Table 4
Variable Means by Submarket for Mixture Model

Variable	Submarket			
	1	2	3	4
Price	308601.85	258124.52	379170.11	320625.43
BUILT	1945.49	1947.64	1943.85	1947.97
FLR	132.84	133.32	148.98	151.84
LANDO	0.30	0.44	0.27	0.19
LAND	762.30	793.13	837.87	865.68
CHATTELS	14,118.65	13025.57	12065.74	9137.29
GOOD	0.13	0.17	0.22	0.25
AVER	0.82	0.74	0.73	0.66
POOR	0.05	0.09	0.06	0.10
T	0.36	0.32	0.36	0.39
N	3396	704	400	212
%N	72.1%	14.9%	8.5%	4.5%

Table 5

Parameter Estimates by Submarket for Spatially Contextual Mixture Model

Variable	$\gamma = 0.065(3.439)$							
	Submarket							
	1		2		3		4	
	Coeff(t)	t statisti	Coeff(t)	t statisti	Coeff(t)	t statisti	Coeff(t)	t statisti
π	0.835	75.838	0.072	8.276	0.064	60.314	0.017	1.864
Constant	3.662	5.595	60.006	4.964	37.843	2.369	39.943	0.521
BUILT	-0.046	-0.541	-7.313	-4.307	-4.085	-1.942	-4.362	-0.417
FLR	0.015	3.057	0.737	7.241	0.906	10.466	1.431	6.212
LANDO	-0.042	-1.872	0.425	0.774	-0.072	-0.077	-3.790	-0.991
LAND	-0.006	-1.815	0.077	0.917	0.021	0.153	-0.584	-1.007
CHATTELS	0.972	226.692	0.028	4.522	0.144	10.441	0.254	10.570
AVER	-0.012	-3.551	-0.407	-5.196	-0.192	-2.559	-0.063	-0.237
POOR	-0.019	-3.821	-0.446	-3.373	-0.663	-2.973	-0.940	-2.929
T	0.006	1.496	0.105	1.088	0.426	2.452	0.477	1.234
σ	0.051	60.922	0.255	10.624	0.300	6.706	0.644	4.589

Table 6

Variable Means by Submarket for Spatial Contextual Mixture Model

Variable	Submarket				
	1	2	3	4	5
Price	292285.03	476500.00	384105.61	315466.53	530912.53
BUILT	1945.73	1945.45	1945.90	1945.00	1955.45
FLR	131.01	176.33	170.99	160.56	130.30
LANDO	0.32	0.24	0.19	0.17	0.30
LAND	763.22	891.82	876.52	855.17	1015.65
CHATTELS	13574.90.65	13508.69	12988.05	9361.66	19000.30
GOOD	0.13	0.38	0.26	0.39	0.30
AVER	0.81	0.59	0.70	0.56	0.64
POOR	0.06	0.03	0.04	0.06	0.06
T	0.35	0.41	0.40	0.45	0.34
N	4281.00	264.00	161.00	36.00	33.00
%N	89.5%	5.6%	3.4%	0.8%	0.7%

References

Abend, K., T.J. Hartley and L.N. Kanal (1965). "Classification of Binary Random Patterns," *IEEE Transactions on Information Theory,* 11, 538-544.

Aigner, D. (1973). "Regression with a Binary Independent Variable Subject to Errors of Observation," *Journal of Econometrics, 1,* 49-60.

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

Bajic, V. (1985). "Housing-Market Segmentation and Demand for Housing Attributes: Some Empirical Findings," *AREUEA Journal,* 13(1), 58-75.

Besag, J.E. (1986). "On the Analysis of Dirty Pictures," *Journal of the Royal Statistical Society B,* 48, 77-92.

Bourassa, S., F. Hamelink, M. Hoesli, B. MacGregor (1997). *Defining Residential Submarkets: Evidence from Sydney and Melbourne,* Real Estate Research Unit Working Paper 3, Department of Property, University of Auckland.

Dale-Johnson, D. (1982). "An Alternative Approach to Housing Market Segmentation Using Hedonic Price Data," *Journal of Urban Economics,* 11, 311-332.

Dempster, A.P., N.M. Laird and D.B. Rubin (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society B,* 39, 1-38.

Efron, B. and R. Tibshirani (1986). "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science, 1,* 54-77.

Everitt, B.S. (1981). "A Monte Carlo Investigation of the Likelihood Ratio Test for the Number of Components in a Mixture of Normal Distributions," *Multivariate Behavioural Research,* 16, 171-180.

Gabriel, S. (1984). "A Note on Housing Market Segmentation in an Israeli Development Town," *Urban Studies,* 21, 189-194.

Goodman, A. (1981). "Housing Submarkets Within Urban Areas: Definitions and . Evidence," *Journal of Regional Science,* 21(2), 175-185.

Grigsby, W., M. Baratz, G. Galster and D. Maclennan (1987). "The Dynamics of Neighbourhood Change and Decline," *Progress in Planning,* 28(1), 1-76.

Hathaway, R.J. (1985). "A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions," *Annals of Statistics,* 13, 795-800.

Hernandez-Avilla, A. (1979). *Problems in ClusterAnalysis,* Unpublished D. Phil. thesis, University of Axford.

Keifer, J. and J. Wolfowitz (1956). "Consistency of the Maximum Likelihood Estimates in the Presence of Infinitely Many Incidental Parameters," *Annals of Mathematical Statistics,* 27, 887-906.

Keifer, N.M. (1978). "Discrete Parameter Variation: Efficient

Estimation of a Switching Regression Model," Econometrica, 46, 427-434.

Kendig, H. (1976). "Cluster Analysis to Classify Residential Areas: A Los Angeles Application," *JAPA,* 42, 286-294.

Kennedy, G. (1995). *A Spatial Analysis of Variation in Rural Real Estate Prices Across Homogeneous Land Market areas in Louisiana,* Unpublished Ph.D. Dissertation, LSU, Baton Rouge, Louisiana.

Lee, L. and R. Porter (1984). "Switching Regression Models with Imperfect Sample Selection Information - With an Application on Cartel Stability," *Econometrica,* 52(2), 391418.

MacLennan, D. and Y. Tu (1996). "Economic Perspectives on the Structure of Local Housing Systems," *Housing Studies,* 11(3), 387-406.

Maddala, G.S. (1983). *Limited-dependent and Qualitiative Variable in Econometrics,* New York: Cambridge University Press.

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

Pace, R K.; R. Barry and C.F. Sirmans (1998). "Spatial Statistics and Real Estate," Journal of Real Estate Finance & Economics, 17, 5-13.

Pace, R K. and O.W. Gilley (1997). "Using the Spatial Configuration of the Data to Improve Estimation," Journal of Real Estate Finance & Economics, 14, 333-40.

Rothenberg, J., G. Galster, R. Butler and J. Pitkin (1991). *The Maze of Urban Housing Markets,* University of Chicago Press, Chicago, IL.

Rosen, S. (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," Journal of Political Economy, 82, 34-55.

Schnare, A. and R. Stryuck (1976). "Segmentation in Housing Markets," *Journal of Urban Economics,* 3, 146- 166.

Smith, C. and M. Kroll (1989). "Utility Theory and Rent Optimization: Utilising Cluster Analysis to Segment Rental Markets," *Journal of Real Estate Research,* 4, 61-71.

Titterington, D.M., A.F.M. Smith and U.E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions,* New York: John Wiley and Sons.

Wolfe, J.H. (1971). "A Monte Carlo Study of the Sampling Distribution of the Likelihood Ratio for Mixtures of Normal Distributions," *Technical Bulletin STB* 72-2, San Diego: U.S. Naval Personnel and Training Research Laboratory.

Footnotes

l A finite mixture model iS also an option when prior sample separation information is available. However, a switching regression model can not be used when there is no information on which to base the choice equations. Maddala (1983) offers a good discussion of the differences between switching regression and mixture models.

2 The parameter $1EM$ can be determined from the constraints (2) as $I\_\sim,mM=1$ $11m$.

3 Each observation, gives rise to a potential singularity on the edge of parameter space. 4 The mean and standard deviation of the LAND are calculated using the properties which have recorded lot sizes. 5 I am not suggesting that the sorting process be modelled. Modelling the sorting process would be inconsistent with the reduced form nature of the basic hedonic model. A theoretically solid sorting model would require the specification of household preferences, not present in the hedonic model. However, a sorting model would be a nice addition to a simultaneous equations model of the demand for housing. 6 Simply adding household characteristics to the specification of the mixing probabilities does not accomplish this because it does not introduce a spatial context.