

Hedonic Regression Models When Unmeasured Quality Differences are Present¹

by

Li Feng
Florida State University

Stefan C. Norrbin*
Florida State University

David W. Rasmussen
Florida State University

and

Jeffrey S. Ueland
Ohio University

This version: September 26, 2005

Abstract:

Hedonic regressions on housing data reveal instabilities that are likely to come from sub-groups of houses in the full sample. Past research has dealt with such sub-groups by *ad hoc* constraints. In this paper we propose a statistical methodology, namely the Mahalanobis distance, to select houses that statistically belong to the same sub-group of houses. We show that this technique is useful in generating coefficient estimates that are more consistent with *a priori* expectations, and also show that this technique can be used to examine whether the results are robust to a potential missing variable bias.

JEL: R21, R15

Keywords: Mahalanobis distance; sample selection; hedonic regression; unobserved quality

¹ A prior version was circulated under the title “Measuring the Capitalized Value of Public Policy and Neighborhood Attributes on House Prices: An Empirical Assessment of Bias in Hedonic Models” and was presented at the 26th Annual Research Conference of the Association for Public Policy Analysis and Management in Atlanta, GA. The authors gratefully acknowledge the support provided by the Federal Home Loan Bank-Atlanta to the Housing Affordability Research Project conducted by the DeVoe L. Moore Center at Florida State University. We gratefully acknowledge the research assistance provided by Moriah Bellinger and comments from Keith Ihlanfeldt, Paul Beaumont, and participants in the DeVoe L. Moore Center seminar.

* Corresponding author. Department of Economics, Florida State University, Tallahassee, FL 32306-2160. Phone: (850) 644-7204, Fax 850-644-4535; E-mail: snorrbin@coss.fsu.edu.

I. Introduction

Hedonic models of housing markets are used to predict prices and to assess the impact of public policies and neighborhood characteristics on dwelling value. These models have been used to evaluate the capitalization into house prices of spatial variations in public services (such as education, transit availability, and public safety), environmental hazards (such as noise and brownfields) and neighborhood attributes (such as race and socio-economic status). This paper provides evidence that standard hedonic models may not yield accurate estimates of the impact of these spatial attributes that affect house value and argues that better estimates of these effects might be generated by estimating hedonic regressions that use observations selected by a generalized distance function that yield more homogenous samples of sold dwellings.

When hedonic models are used to estimate the impact of spatial characteristics on house prices, the issue of sub-markets has received considerable attention. Dwellings in a sub-market of a metropolitan area are defined by Bourassa, et al. (1999) to be reasonably close substitutes for one another and imperfect substitutes for units in other sub-markets. Goodman and Thibodeau (2003) point out that market segmentation is typically an *ad hoc* enterprise; markets can be stratified by race, income, structure type, and neighborhood characteristics. But in practice, sub-markets often are spatially defined in terms of census tracts, zip codes, neighborhood quality (nested models), and aggregations of block group data constructed using geographic information systems. Goodman and Thibodeau conclude that “smaller is better” (p. 20), which is no doubt true if smaller implies a more homogeneous neighborhood.

Straszheim’s (1975, p. 28) observation that “variation in housing characteristics and prices by location is a fundamental characteristic of the urban housing market” suggests that sub-markets need not be contiguous. Holding socio-economic status and neighborhood characteristics constant in two non-adjacent census tract “sub-markets,” a high quality home in one census tract may be a close substitute for high quality homes in the other while lower priced homes in the same tract are highly imperfect substitutes. In fact, smaller may be better for defining market segmentation because it may group dwellings of similar vintage and construction, thereby

controlling for some unmeasured quality differences. With our apologies to Gertrude Stein who asked if a rose is a rose, it is obvious that all square feet are not identical. Relatively high priced homes will have many unmeasured aspects of quality such as finer decorations and fixtures, floor coverings and landscaping so that attribute coefficients in a hedonic regression in this market might be quite different from a large sub-market of homes that has more homogeneous characteristics. Dwellings at the low end of the market may also be characterized by unmeasured quality differentials.

This paper examines different methods of selecting a sample of houses that is less likely to be affected by unmeasured quality differentials. We discuss various ad hoc methods used in prior research to select a relatively homogeneous sample of dwellings. The ad hoc nature of the restrictions usually used imply that each investigator may come to a different conclusion about the value of public policy and neighborhood attributes depending on how the sample is selected. This problem is not amenable to an obvious solution and only repeated experiments using data from various markets and alternative sample selection criteria will illuminate the “true” impact of neighborhood characteristics on house prices.

We explore the potential advantages of using a statistical technique that can select a homogenous group of “normal” houses. We find that using a generalized distance function to select a sample that minimizes unmeasured quality variations leads to a hedonic regression with plausible and stable parameters. This process has four advantages. Firstly, the technique is statistically based, thus the researchers’ priors are not a factor in filtering the data. Secondly, the technique allows for covariation among variables. The usual ad hoc methods are focused on setting constraints for each variable individually. In contrast, the Mahalanobis distance function selects the data according to both variation and covariation of the explanatory variables in the system. Thirdly, the technique is robust to multicollinearity and has a simple statistical interpretation. Finally, the technique provides a simple way to assess the potential size of the problem associated with the unmeasured quality variables.

The following section presents a hedonic regression for the Duval County housing market. The results indicate that several parameters have suspect coefficients and some have signs that are inconsistent with theoretical priors. The third section presents a brief survey of the ad hoc methods that have been used to determine which observations are used in the analysis. We examine hedonic regressions using *ad hoc* methods that might adequately purge observations with unmeasured quality attributes. These results suggest that the high priced sub-market is composed of less homogeneous attributes and that the valuation of non-housing attributes is not constant across these sub-markets. In section four we explore the efficacy of using a generalized distance function developed by Mahalanobis (1936) to identify samples of relatively homogeneous homes in Duval County that provide hedonic estimates that differ in important ways from models using the full sample. Conclusions are presented in section five.

II. An Empirical Assessment of the Value of Housing Characteristics

We conduct an empirical assessment of the values of characteristics using tax roll data for detached single family dwelling sales during 1995 in Jacksonville, FL.² These data are particularly useful, because they cover a single jurisdiction, Duval county.³ We identified 7,645 single-family detached dwellings that were “qualified” as arm’s length market transactions according to the County Assessor, and were likely to be market transactions.⁴ A semi-log

² Two other potential data sets, namely the American Housing Survey and the Multiple Listing Service data were considered to have too many drawbacks. The American Housing Survey has drawbacks in that the house value is reported by the occupant, the number of square feet in the dwelling is unavailable, and there is lack of objective neighborhood information. Multiple Listing Service (MLS) also has drawbacks in that it has data that are proprietary and may account for a relatively small percentage of housing market transactions.

³ See Lynch and Rasmussen (2004) for a discussion of the ways in which these data are likely to be useful for hedonic regression estimates.

⁴ “Qualified” sales according to the County Assessors sometimes involve some questionable procedures, such as arm’s length determination “by telephone.” Therefore we eliminated some sales from the County Assessor’s “qualified” sales according to the following criteria. Because the parcel can be sold before a structure is built, we excluded all sales of houses less than a year old. Furthermore, nine observations were eliminated because we could not verify lot size information using parcel data available from the tax assessor. Finally, after initial regressions were run we eliminated 20 observations for which the absolute value of the residual exceeded five standard deviations, for example one of the houses was reported to have three bedrooms and twenty bathrooms, whereas another dwelling was reported to have a 20,000 acre lot.

regression methodology is selected following the prior literature.⁵ The empirical specification is given by:

$$LPrice_h = \alpha + \sum_{i=1}^I \beta_i X_{ih} + \sum_{j=1}^J \delta_j Y_{jh} + \sum_{k=1}^K \gamma_k Z_{kh} + \varepsilon_h ,$$

where the dependent variable, $LPrice_h$ is log of the sales price for the h th house, β_i is the regression coefficient for the i th dwelling characteristic X_{ih} , δ_j is the regression coefficient for the j th neighborhood characteristic Y_{jh} , and γ_k is the regression coefficient for the k th location characteristic Z_{kh} , and ε_h is an error term that may be spatially correlated. A complete list of these variables and summary statistics is presented in Appendix A.

The dwelling characteristics are variables associated with the house construction or the lot that the house is built upon. The number of square feet (SQFT) measures the size of the house. A larger square foot size would increase the cost of building the house and would thus be expected to increase the selling price. In addition some factors pertaining to the construction are expected to increase the price, namely: number of bathrooms (BATHROOMS), number of bedrooms (BEDROOMS), central air (CENAIR) and heat (CENHEAT), and the existence of a fireplace (FIREPLACE). The log of the age of the house (LAGE) measures the depreciation of the dwelling and would thus be expected to lower the price. The number of acres, LOTSIZE, is expected to increase the price of a house with a larger lot. Additional features such as existence of a pool (POOL) and the presence of a garage (PARKING) should increase the price of a house.

Unique neighborhood characteristics for each house are defined using a geographic information system. Using block group data from the 2000 decennial census and employing the ARCGIS system, information on the economic and demographic characteristics of the area immediately surrounding each observation were collected. To gather neighborhood data that are unique to each observation in the sample, the latitude and longitude of each house in the sample is found using a geographic coding program. A radial distance of one-half mile is swept around

each observation to generate the neighborhood characteristics that are unique to each dwelling.⁶ Because the information is only available at the block group level, these data are retrieved via "proportional grabs." Under this approach the neighborhood includes all census block groups that are entirely within the circle as well as those that are partially included. In effect, it is assumed that household characteristics are distributed evenly throughout the block groups, so characteristics of block groups that are partially in the circle are also included in the estimation of neighborhood characteristics.⁷

Seven variables are created in this way: population density (POP DENSITY); percentage of households that are homeowners (OWNER%); average household income (AVERAGE INCOME); percent black (BLACK%); percent Hispanic (HISPANIC%); percent of head of households over age 50 (OVER50%) and percent white collar workers (WHITE COLLAR%). Population density is assumed to be an inferior good, while owner occupied single-family dwellings are presumably better maintained and multi-family units are presumed to generate negative externalities. Lower mobility among older households suggests a lower supply of houses for sale in neighborhoods with a relatively large population over age 50. Minority population, social class, and income are pertinent determinants of house value for the three reasons. First, because all dwellings are located within a single Duval County school district, differences in school quality are likely to be the product of the socio-economic characteristics of the school's catchment area. Assuming that house values, in this context, are determined by dwelling and lot characteristics, neighborhood attributes, and school quality, and that school quality is determined by socio-economic characteristics of the population it serves, ours is essentially a reduced form model.⁸ Second, perceptions of public safety may be higher in affluent

⁶ Lynch and Rasmussen (2004) show that hedonic coefficients are relatively stable for neighborhoods defined from .1 to .5 miles surrounding a dwelling using MLS data for Duval County.

⁷ These variables are from the 2000 Census that seemed preferred to the 1990 data in that the more recent data would capture whatever expectations generated by socio-economic trends in the neighborhood early in the decade. We use 1995 sales data in this paper because we intended to focus our work on comparing the empirical results generated by tax data to those generated by MLS data we have for 1995. The essential findings reported below are similar between the two data sets.

⁸ For a more extensive discussion of the impact of variable school catchment areas within a single school district, see Lynch and Rasmussen (2004).

communities and many other aspects of community life are probably perceived to be superior in such neighborhoods. Among these advantages are superior parks and recreation facilities, better shopping opportunities, more aesthetic appeal, and other intangibles of neighborhood quality. Finally, encroachment of low-income households may trigger fears of diminished neighborhood quality and a corresponding decline in property values.⁹

Four location variables are included to augment these neighborhood characteristics: distance (measured in miles) to the center of the central city DIST_CBD, distance to the St. Johns River DIST_STJOHNS, and distance to the Atlantic Ocean, DIST_ATLANTIC. Sales prices should be negatively correlated with distance to such employment centers and amenities. Finally, the variable that captures water front location, WATERFRONT, should capture the positive value of having a house located immediately on a type of water, such as ocean, lake or river.¹⁰

Table 1 reports the results for a hedonic regression using the full sample of 7,645 dwellings. The dependent variable and age were logged, whereas the other variables are in levels. Column 1 in Table 1 reports the estimates for the OLS version, whereas the second column 2 shows the estimates corrected for spatial autocorrelation. The results indicate that such a correction did not lead to a major change in the results.¹¹ The results are generally satisfactory, with a coefficient of determination of almost .80. The table shows that most of the dwelling characteristics are significant and have the expected sign, except for central heat that has an insignificant coefficient. Probably this is due to the small number of houses that do not have central heat (see appendix A). The lot size variable has the expected sign, but is very small. An acre of land is only valued at \$8,605 for the average house. The average house has a lot size of 0.31 acres, so these estimates suggest the cost of the lot accounts for only a small fraction of the sales price.

⁹ The role of race in this process is somewhat ambiguous. Leven et al. (1976) indicated that lower socio-economic status, not race per se, is the trigger that prompts fear of declining property values. However, race per se is likely to have an impact to the extent that homeowners engage in statistical discrimination, i.e., assume that an increasing number of black households in an area implies a decline in socio-economic status.

¹⁰ The definition and sample means of the variables are presented in Appendix A.

¹¹ In this paper we chose to correct the spatial autocorrelation using a spatial lag model. A single lag model with an area with a 0.5 mile radius is used for all regression. We also tried correcting using a spatial error model, however, in most cases the choice of a correction did not alter the conclusions. See Anselin (1999) for a discussion of spatial error versus spatial lag models.

Neighborhood characteristics generally have the expected sign. Population density has the expected negative sign and prices are higher in neighborhoods that are more affluent, have fewer blacks, a higher percentage of older residents, and more white-collar workers. Sales prices are higher when there are more owner occupants in the neighborhood, but this effect diminishes as this proportion rises since the squared term is negative and significant. The effect on the house price of the Hispanic population variable is positive and significant at the 90% level. This finding is contrary to expectations.

Each of the location variables is statistically significant. However, contrary to standard urban theory, dwellings that are further from the central business district, *ceteris paribus*, have higher prices, and dwellings further away from St. John's river have a higher price. Thus Table 1 indicates that most of the variables have reasonable estimates, but that puzzling magnitudes and signs exist on a few of the coefficients. Such odd estimates might be an indication of a heterogeneous sample that is characterized by unmeasured quality differences, especially at the extremes of the price distribution. Thus some technique of allowing for such unmeasured quality differences may be necessary. We turn to this task in the next section.

III. Missing Variable Bias

Most papers assume that one can differentiate the effects of the observable variables and unobservable variables, treating the effects of unobservable variables as orthogonal to the observable variables. Unfortunately the housing market is a relatively heterogeneous market and many omitted variables can affect the estimates of the observable variables. For example, the existence of parks, sidewalks, landscaping, trees, lakes, and rivers can bias the value placed on observable characteristics of houses. The simplest solution is to find proxies for the unobservable variables. However, such a solution is often not possible in practice as many of the unobservable variables are difficult to proxy, and adding a large number of variables will result in a high degree of multicollinearity. In the above regression we add many neighborhood variables to eliminate as

much as possible of the unobservable variable bias, but we acknowledge the fact that some missing variables remain.

Three solutions to the missing variable problem have been considered, namely: binary control variables, mechanical adjustments, and data groupings. One can use binary control variables to try to find proxies for the omitted neighborhood variables. An interesting attempt to control for neighborhood effects is Black (1999), who estimates the effects of education as it is reflected in the housing prices. Noting that all neighborhood variables cannot be proxied with observable variables, she adds binary variables for each overlapping school district. Thus if houses are close enough to a school zone border, then they are grouped together. This is analogous to a fixed effects model, where one expects the fixed effects to be common to houses in close proximity to each other. Black (1999) finds the value of schools to be cut in half when these binary values are included, indicating a substantial effect of the missing variables.¹²

Secondly, one could use mechanical corrections. Two such mechanical corrections are commonly used, namely functional form searches and spatial error corrections. The traditional functional form transformations, such as semi-log or double-log transformations might better fit the function, by allowing the observable variables to proxy the omitted variables better. In addition spatial error correction methods could improve the estimated coefficients, as the omitted variables from spatially close houses may exhibit similar errors and thus a variance-covariance adjustment may provide some correction for the omitted variable problem. In this paper we try different types of functional forms, as well as control for spatial errors.

Thirdly, the data could be grouped to minimize the chance of omitted variable effects. If we can select data that belong to a “normal” group then the data are less likely to be affected by an omitted variable bias. In the literature several methods have been used to create a relatively homogenous data set. All studies attempt to eliminate outliers or actual mistakes in the data. Some also eliminate observations at the extremes of the price or size distribution, which may

¹² Because Duval County is one jurisdiction, there are not any fixed borders that systematically affect the housing prices.

have the effect of reducing unmeasured quality differences in the sample. These practices are implicit attempts to construct a homogenous sub-sample by filtering the full data set and, in some cases, eliminate a significant fraction of the data. For example, Gatzlaff and Haurin (1997) estimate that they reject 10 percent of the observations, Case and Schiller (1987) reject 14.5 percent, and 10 percent are eliminated in Meese and Wallace (1997), although the latter do not explain the criteria they used to reject the observations. The methods focus on determining “normal” expected values for individual variables, and excluding data where the values are outside of the “normal” interval. Among the criteria used are: price/assessed value ratio, lot size, house size and price of the house.

There are many possible types of groupings. We divide the groupings into ad hoc and statistical groupings. An ad hoc type grouping is commonly used in housing papers. Implicitly cutting certain data from the regression is a type of grouping. The “outlier” methods implicitly divide the data into a “normal” and an “extreme” group. Two methods are used, namely a dependent variable and an independent variable ad hoc selection. In many papers the dependent variable is implicitly grouped into a “normal” house and a “unique” house by price limits. A number of studies eliminate dwellings according to a price criterion. It should be noted that this may be a problematic way of selecting a homogenous group, since this might lead to a truncation bias due to the fact that the data are truncated using the dependent variable. Some authors eliminate houses with low prices only, whereas others eliminate both houses with low and high prices. Figlio and Lucas (2004) eliminate all sales of developed property below \$10,000. Boarnet and Chalermpong, (2001) eliminate dwellings that sold for less than \$25,000 and Gatzlaff and Haurin (1997) include only those selling for more than \$8000. Dale-Johnson and Yim (1990) eliminate dwellings sold for less than \$15,000 and over \$350,000 while Delaney and Smith (1989) choose a narrower range of those with a selling price between \$20,000 and \$175,000. Haurin and Brasington (1996) also use a price criterion and eliminated houses with transactions prices over \$400,000 or below \$10,000. We use this approach in this paper to show the dangers of censoring the dependent variable.

The independent variables could also be censored to keep only “plausible” values. A large number of studies have eliminated dwellings based on a square foot criterion of either the structure or the lot size. Dale-Johnson and Yim (1990) limit their analysis to dwellings with between 500 and 4000 square feet while Gatzlaff and Ling (1994) limit their sample to units with between 800 and 6000 square feet. Dale-Johnson and Yim, (1990) also require lots to have more than 2000 square feet but less than 250,000, and Haurin and Brasington (1996) eliminate lot sizes greater than two acres. This method has been applied in an ad hoc manner, and has ignored the possible covariation between the independent variables. Therefore we turn to statistical methods that allow for the covariation between the independent variables.

Three statistical methods have been used in the economics literature. The first is the propensity score technique to design a control group. This is often used when subsets of data have characteristics that are not likely to be shared with the whole data set. In the case of the hedonic model, there are no priors about what characteristics the “normal” group should have, therefore this method is not useful in our study. Another statistical method is cluster analysis. Bourassa et al. (1999) and Abraham et al. (1994) use a cluster analysis approach to group housing data. Unfortunately cluster analysis is sensitive to large numbers of estimated coefficients, due to a sensitivity to multicollinearity.¹³ The third type of statistical approach is a general distance function. This method selects data that are statistically close to the centroid of the variance-covariance matrix of all variables. This statistical method has two important advantages: it is robust to multicollinearity, and it is easy to use because the distances are in standard deviations. Thus the 95 percent confidence interval is easy to construct.

IV. Censoring the Dependent variable

Before turning to our statistical method we present a case where an *ad hoc* method has been used to divide the data into three categories based on the price. Our operating assumption is that

¹³ Bourassa et al. (1999) use factor analysis to reduce the multicollinearity, and subsequently use cluster analysis on the factors. This two step process would be interesting to compare to the Mahalanobis distance approach in future research.

the low and high priced dwellings will have substantial unmeasured quality differentials and the regressions will consequently account for a smaller portion of the variation in sales price. It should be noted that such a restriction on the variability of the dependent variable is problematic. Such constrained regressions may suffer from a truncated regression bias. Furthermore, recent work, for example Bollinger et al. (2005) have shown that the removal of any data from the dependent variable might lead to a bias. Thus they advocate not bounding the dependent variable, and in fact argue that no outliers should be removed either. We acknowledge such concerns, and merely show the effect of ad hoc practices common in the past literature.

Table 2 reports the regression results of the *ad hoc* method of selecting a sub-group. The first regression includes only low priced dwellings, those with a recorded sale price of under \$42,000 which account for 10 percent of the sample. Regression 2 is limited to observations that are in the middle 80 percent of the price distribution that includes dwellings that sold from \$42,000 to \$145,000. Only those homes selling for more than \$145,000 are included in regression 3. Thus regression two eliminates more observations than any of the studies reported above.

It is apparent that the variables included here account for a much larger share of price variation in the middle 80 percent of the sales distribution ($R^2 = .76$) compared to the variation in prices of homes selling for under \$42,000 (adjusted $R^2 = .31$). Similarly, unmeasured quality attributes appear to play an important role in the more expensive homes relative to the middle 80 percent of the price distribution, as reflected in the lower explanatory power of the model ($R^2 = .45$). The regression that is limited to the 10th through 90th percentiles of the sales price distribution yields more accurate predictions of price than the regressions in Table 1. The root mean square error (RMSE), is .1537 in the 80 percent sample versus .2424 in the full sample. The weighted sum of the RMSE for the three regressions in Table 2 is .1815, which is substantially less than the full sample RMSE. The summary statistics of the regressions presented in Tables 1

and 2 indicate that the predictive power of hedonic regressions can be somewhat improved by limiting the sample to dwellings that arguably have more homogeneous characteristics.¹⁴

However, the improved predictive power of the 80% sample does not solve the puzzling coefficients from Table 1. The coefficient on LOTSIZE actually decreases to 0.071, implying an even lower value of additional acres of the lot size. The distance measures from the central business district and St. John's river remain positive, and increase substantially in size. The only puzzling coefficient that is resolved is from the HISPANIC variable. This value becomes significantly negative in the 80% sample. The positive HISPANIC coefficient in Table 1 appears to come from some high priced houses, as this variable has a significant positive coefficient in the top 90-100% sample in Table 2. A few other coefficients show substantial changes between Table 1 and the 80% sample in Table 2. Central air has a substantially lower coefficient in Table 2. In fact, the value is only a third of the fairly reasonable value in Table 1. The waterfront variable also falls sharply, presumably because the expensive houses on Jacksonville beach are excluded from the 80% regression.

Hedonic regression analysis often plays an important role in public policy analysis to discover how households value neighborhood and environmental attributes. Access to public transportation, school quality, noise, and public safety are among the aspects of urban life that have been examined in hedonic models, as have the effect of minority populations on house values. If the income elasticity of demand for some neighborhood attribute is positive, the estimated coefficient using the entire data set may not accurately reflect what the "average" household is willing to pay. Consider the impact of BLACK% on property values in Tables 1 and 2, a variable that is significantly negative in every regression. When the entire data set is used in regression 1, a one percentage point increase in the percent black in the neighborhood lowers

¹⁴ Since the model explains a much larger portion of the variation in price in the 90-100 percentile regression relative to that using the 0-10 percentile of the data, it appears that a more scientific method of determining a homogeneous sample would reduce the number of low dwellings priced at the low end of the distribution. As is shown below, it appears that for our data 70 percent of the dwellings sold is the upper bound for a sample that can be viewed as homogeneous. It should be noted that the improvements in explanatory reported here do not provide a compelling reason to select more homogenous samples of sales because altering the functional form can easily yield greater predictive power when the goal is simply to predict price.

property values by .0033 percent. When the sample is limited to the 10th to 90th percentiles in regression 2, the coefficient is significant but is only -.0007; the implication being that the estimate of race on house values in Table 1 could be wrong by a factor of five. The market wide estimate of race could be affected by the high price segment where relatively few blacks reside.

These results suggest that it may be unwise to rely on the veracity of hedonic models that use data from the population of sold dwellings to estimate how households value neighborhood amenities. Perhaps more accurate price estimates can be obtained when the market is segmented to be more homogeneous. The results in Table 2 support the notion that unmeasured quality differences compromise efforts to estimate prices at the low end and high end of the market. However, it is questionable whether this type of sub-sample selection, using the price as a criterion, is appropriate. We now explore a statistical method that allows us to minimize unmeasured quality differences by generating a population of similar dwellings. We can then evaluate the efficacy of alternative sample selection methods when using hedonic regressions to estimate the impact of neighborhood and location variables on house prices.

V. Statistical Selection of a Homogeneous Housing Market

A method that selects data on the basis of both variation and covariation was developed by Mahalanobis (1936). This generalized distance function provides a way to identify a “standard” house and measures how similar other dwellings are to this benchmark.¹⁵ The Mahalanobis’ distance function is defined as:

$$D^2 = (X - M_x)' \Sigma_x^{-1} (X - M_x) \quad ,$$

Where D is the Mahalanobis distance, X are the house characteristics, M_x are the means of the house characteristics for the data set, and Σ_x is the variance-covariance matrix for all of the characteristics. The Mahalanobis distance is designed to create a concept of distance between data points when the relevant variables have different units of measurement and standard

¹⁵ See Klecka (1980) for a description of this method.

deviations. This method calculates the minimum distance from a “standard” house to create a sample of sold dwellings that are most like the reference dwelling. Note that the characteristics can receive very different weights according to how common the units observed are. For example, if a dummy variable has a unit value in only 5% of the data, then this variable is likely to receive a large weight as most of the data has a zero value.¹⁶

The “standard” dwelling is the house that closest resembles the mid-point of the covariance matrix of all the variables. Thus, all the variables are considered jointly to find a house that most closely resembles a “standard” house in all dimensions. Note that such a “standard” house does not exist, but would be a house that is average in all dimensions. In application the “standard” house would be the house with the lowest Mahalanobis score. Then all other houses can be ranked in the order that they resemble the “standard” house. This method thus allows us to rank houses in all dimensions jointly, instead of examining each dimension separately as is done in most studies. Using the ranking one can then eliminate houses that are not “close enough” to the “standard” house using a statistical confidence interval.

To illustrate how the Mahalanobis distance deals with the problem of unmeasured quality differentials we generate Monte Carlo data using the following three different processes. Label the first group as the “normal” quality group because it is the most common type of house, the second group as the “high quality” dwelling, and the third as the “low quality” dwelling. The difference between the groups is the fact that the second and third groups have an unmeasured quality variable that the researcher is unaware of. We generate the first group, the “normal” group according to the following equation:

$$P_h = \alpha + \beta X_h + \gamma Z_h + \varepsilon_h ,$$

¹⁶ This method has frequently been used to identify outliers. See, for example, Rocke and Woodruff (1996) who argue that Mahalanobis distance is a useful method to identify outliers, especially when these outliers are not clustered. Furthermore the method has been used in other scientific fields to identify sub-groups. For example Dunn and Duncan (2000) use Mahalanobis in an example where habitat suitability needs to be established to determine if one would expect the occurrence of species in a particular habitat.

where X_h , Z_h and ε_h are random variables with a zero mean and unit variance. The second and third groups include an unmeasured quality variable:

$$P_h = \alpha + \beta X_h + \gamma Z_h + \lambda Y_h + \varepsilon_h$$

where the Y_h variable is a measure of the unobserved quality and λ is the effect of the variable on the sale price of the house. Assume, for simplicity, that the Y_h and the Z_h variables are perfectly correlated.¹⁷ The price of the house, P_h , is then constructed by adding the weighted values of the three characteristics X_h, Z_h, Y_h , and the specific idiosyncratic shock of ε_h . Further assume that all groups have an $\alpha = 100$, $\beta=10$ and $\gamma=-30$, and that the coefficient on Y_h differs for the three groups. For the first group (“normal”) the $\lambda = 0$, but in the second group (“high quality”) $\lambda = 90$, and in the third group (“low quality”) the $\lambda = -30$.

In terms of this example, one can think of the Monte Carlo data as houses that have two attributes. The first is the number of square feet, and the second is the distance to the central business district; these attributes determine the price of the house. If the house is located far away from the central business district and is large, then the house is of a different group from the “normal” house. Similarly if the house is small and located in the central business district area then it is different from the “normal” house. Thus a subset of the houses have a nonzero value for the unobservable characteristics variable.

A sample of 7,645 (the same number of observations as in the above data for Jacksonville) is used for the Monte Carlo data, and 1000 iterations are performed to make sure that the random draws are representative. The results show a dramatic bias even when only 12.50% of the data

¹⁷ The degree of correlation impacts the exact bias of the coefficient. To simplify the exposition we report results with perfect correlation. This amounts to making this a threshold model, where for a certain level of both variables the value of the coefficient on the Z variables switches.

points are in the “high quality” or “low quality” groups.¹⁸ The results for the 1000th iteration of the exercise are:

$$P_h = \begin{matrix} 112.25 & + & 15.89 X_h & - & 21.83 Z_h \\ (267.87) & & (32.84) & & (-44.53) \end{matrix}$$

with t-statistics in parentheses.¹⁹ The γ coefficient is substantially less than the expected value of -30.00. The existence of an omitted variable results in a bias of 26% for the γ coefficient. The existence of two groups with unobserved quality variables, also causes the coefficient on both the constant and the X_h to be biased. The β coefficient has an upward bias of almost 63%. Thus the existence of sub-groups in the data can cause the hedonic regression to estimate an incorrect magnitude and sign for the coefficients.

Applying the Mahalanobis method to these Monte-Carlo data leads to the results presented in Table 3. The left column shows the cutoff used for the data. If the full sample is used then no cutoff is used. In contrast if we select out a portion of the data then different levels of cutoffs are used. A cutoff of 4.00 implies that all data within a four standard deviation level of the normal house is used, whereas the 1.00 cutoff implies a much stricter criterion with all data selected being within one standard deviation of the “normal” house. The covariance term refers to the covariance between the regressors X_h and Z_h . The results show that the results change substantially as more data are thrown out. This implies that the omitted variable seriously biases the results. Of course this is the case because the data were constructed in such a way that there is a strong bias from the missing variable, and the Mahalanobis results confirm this. For example, with a covariation of 0.50 the bias of the β -coefficient is 62.96% for the full sample, whereas the bias decreases to 12.99 with a cutoff of 2.00 and is virtually eliminated with a 1.00 cutoff. A similar pattern exists for the remaining covariance specifications. The Monte Carlo data illustrate that the results are highly sensitive to sample selection. Thus a hedonic regression

¹⁸ We use a value of 1.0 for both of the variables, X_h and Z_h , jointly as a cutoff for classifying the observation in the “high quality” group, and a value of -1.0 to classify the observation into the “low quality” group.

¹⁹ This is the zero case in Table 3 with a 0.5 covariance between regressors.

with changing values for the coefficients as more data are added, may be due to an omitted variable bias.

Applying the Mahalanobis method to the actual housing data leads to the estimated Mahalanobis values in Figure 1.²⁰ As expected, adding more observations increases the heterogeneity of the sample. This effect is linear for most of the sample, but the adjusted distance increases exponentially when about 70 percent of the data, 5,352 observations, are included.

To examine the stability of regression coefficients among samples, we use the three sub-samples that are suggested by Figure 1, namely a sample of 10% (n=765), 46% (n=3,581), and 70% (n=5,352), corresponding to one, two and three standard deviations from the “normal house”. Comparing predictive power of regressions in Table 3 with those of Table 1, shown in the last column of the Table, reveals that the explanatory power (R^2) of these models are comparable but the predictive power is higher when more homogenous samples are used. Relative to the RMSE of .2424, in Table 1, for the full sample, the sub-samples in Table 3 all have a lower RMSE. The smallest most homogenous sample has the lowest RMSE at .1159, whereas the less homogenous larger samples have a RMSE of .1730 and .1906 respectively. When we select a sub-group of 10%, we are implicitly assuming that the other sub-group is the remaining 90% of the data. The total RMSE for both groups can be computed to compare this to the full sample predictive power. The total for both groups for the 10% selection is 0.2415, which is slightly below the full sample RMSE. The same computation for the 46% and 70% also generates smaller RMSE of 0.2412 and 0.2388 respectively. Thus the RMSE is reduced slightly. In comparison the ad hoc division according to price in Table 2 resulted in a much lower RMSE than this statistical grouping. However, keep in mind that the reduction in the RMSE in Table 2 was expected as grouping according to the dependent variable should result in a reduction of the RMSE, but also may lead to a truncation bias in the results. According to Bollinger and Chandra (2005) this bias exceeds any improvement in the precision of the estimates due to the decreased

²⁰ In fact the reported Mahalanobis values are adjusted to normalize the values on the house closest to the origin. Thus a house that has a score of 2.0 is two standard deviations away from the house that is most like a “standard” house.

variance of the dependent variable. Their findings are confirmed by our results using the truncated data.

In this statistical grouping we are only concerned with the “normal” group. The Mahalanobis distance sorts all houses that have some unique characteristic or some unusual combination of characteristics. Thus the remaining group potentially could be composed of several sub-groups. For example, the second group can be composed of both large and small houses. Both might be equally far statistically from the “normal” house, and therefore have the same Mahalanobis score, but they would also be very far statistically from each other. If one were interested in analyzing such sub-groups within the discarded group, a better technique would be a cluster type analysis.²¹ In this paper, however, we are only interested in the “normal” group to compute the value placed on each characteristic, for house types that are commonly found in the data set.

Comparing the coefficients reported in Table 4 gives an indication of the degree to which the willingness to pay for specific attributes is sensitive to how the sample is constructed. Focusing on the four puzzling coefficients in Table 1 we can see that three of the four coefficients have the expected effects in Table 4. The value of increasing lot size is seriously distorted by using the entire sample since this coefficient suggests an additional acre only increases sales price by 11.2 percent. When the most homogeneous sample is used this coefficient more than doubles to 24.6 and is substantially larger in the other models as well. The Mahalanobis method provides for a much more homogenous sample in terms of lot size: when 10 percent of the data are used the maximum lot size is .73 acres, when 70 percent is used it is 1.55 and the maximum value of acres in the total sample is 11.73. Lot size as measured by acres has a much larger coefficient in the homogeneous sample that limits lot size to smaller than one acre.²² Modest increments to relatively small lots are highly valued, according this result, an insight that is not apparent in the results reported in Tables 1 and 2. The Lot Size coefficient in column 1 of Table 3 suggests that increase lot size from one-quarter to one-half acre will increase the price of a \$100,000 home by

²¹ See, for example, the use of cluster analysis to construct sub-markets in Bourassa et al. (1999).

²² This outcome is particularly pertinent for Duval County because there are large areas of low density development within this jurisdiction.

about \$5,700. It seems clear that using the entire sample the coefficients estimated are contaminated by unmeasured market characteristics from differences among market segments.

Using the entire sample to estimate the impact of neighborhood and location variables may also yield distorted results. Percent Hispanic in the neighborhood yield a positive and significant sign in the full sample but is negative and significant in all of the regressions that are limited to more homogeneous samples. Estimates of percent black in a neighborhood in Table 2 where much smaller than those reported in Table 1, a result that is not confirmed when the homogeneous samples are used in Table 3. The location variables are also more consistent with *a priori* expectations. In contrast to Table 1, the distance to the central business district has a negative effect in Table 3. Thus a premium is paid for houses that have a small commuting distance to the center of Jacksonville. The distance to St. John's river remains positive in both of the larger samples in Table 3. This suggests that people pay a premium for being further away from the river. This unexpected coefficient may have to do with the location of the river in Jacksonville. The river winds through the city, and the amenities associated with the river vary greatly as some areas are industrial, an attribute that apparently dominates in this sample.

Three of the four puzzling coefficients found in Table 1 are resolved by the results in Table 3, whereas only one of the four coefficients is resolved by the *ad hoc* restrictions in Table 2. Furthermore most of the other coefficients appear consistent with the full sample, in contrast to the *ad hoc* restrictions that lead to a few anomalous coefficient values. This appears to suggest that using the Mahalanobis distance to generate relatively homogeneous sub-samples may be superior to using the full sample or the arbitrary culling of the tails of the price distribution.

V. Conclusions

A significant literature has shown that predicting the sales price of a dwelling is a relatively straightforward process that can be accomplished with standard hedonic models (Goodman and Thibodeau, 2003). Maximizing the predictive power of these models usually requires tinkering with the functional form and, appropriately, do not put a premium on the

veracity of specific coefficients. However, as Schnare and Stryuk (1976) noted, hedonic models are regularly used to estimate the impact of specific neighborhood characteristics, the value of public services, and environmental quality. We have shown that sample selection can have an important impact on the size and sign of coefficients in a hedonic model and that using the entire sample of homes may yield inaccurate coefficient estimates because of substantial unmeasured differences in quality among dwellings.

Using ad hoc methods to eliminate extremes in the data, a relatively common practice as shown in our literature survey, is apparently not an adequate solution to this problem according to the results reported here. Following the past literature we present an ad hoc market segmentation that limits the sample to dwellings selling between \$42,000 and \$145,000s; the middle 80 percent of the sales price distribution. Implausible coefficients in terms of magnitude and theory remain. We show that using a statistically selected homogenous sample using the method developed by Mahalanobis (1936) yields coefficients that appear to be more reasonable in magnitude and are consistent with a priori expectations and theory. In this experiment using data from Jacksonville, FL, we find that about 30 percent of the sample should be eliminated to reduce the problem of unmeasured quality differentials. In addition the Monte-Carlo analysis shows that changing coefficients at different cutoffs is an indication of substantial omitted variable problems. Thus the Mahalanobis method is a viable way to detect if the sample is affected by a substantial missing variable problem.

When using hedonic models to evaluate the benefits and costs of public policy alternatives, these preliminary results suggest that it is prudent to follow Leamer's (1983) admonition that we should be skeptical of specific regression results and be suspicious of the sign and magnitude of coefficients when they differ across specifications and methods of selecting observations. The results also show that the Mahalanobis method provides a simple way of detecting omitted variable effects, by sorting data into groups. If the results are relatively robust to subsets of data then the full sample is unlikely to suffer from an omitted variable bias. More experiments is

useful to evaluate whether the Mahalanobis method is a reliable tool in evaluating the importance of sample selection in hedonic analysis of how house prices are affected by variables such as school performance, crime, and environmental quality. In any event, the usual methods of using hedonic regressions to estimate the value of important social and policy variables by estimating their impact on house prices may need to be re-examined.

References

- Abraham, J. M., W. N. Goetzmann, and S. M. Wachter, S.M., "Homogenous Groupings of Metropolitan Housing Markets," *Journal of Housing Economics*, vol. 3, no.3, p. 186-206, 1994.
- Anselin, Luc, "Spatial Econometrics," working paper, University of Texas at Dallas, April, 1999.
- Anas, Alex and Sung Jick Eum. "Hedonic Analysis of a Housing Market in Disequilibrium", *Journal of Urban Economics*, Vol. 15, p. 87-106, 1984.
- Boarnet, Marlon and Sadsith Chalermpong. "New Highways, House Prices, and Urban Development: A Case Study of Toll Roads in Orange County, CA", *Housing Policy Debate*, Vol. 12, p. 575-605, 2001.
- Black, Sandra, "Do Better Schools Matter? Parental Valuation of Elementary Education," *The Quarterly Journal of Economics*, 114(2), 577-599, 1999.
- Bollinger, Christopher R. and Amitabh Chandra, "Iatrogenic Specification Error: Cleaning Data can Exacerbate Measurement Error Bias," *Journal of Labor Economics*, April, Vol. 23, no. 2, pp. 235-257, 2005.
- Bourassa, Steven, C., Foort Hamelink, Martin Hoesli, and Bryan D. MacGregor, "Defining Housing Submarkets", *Journal of Housing Economics* 8, 160-183, 1999.
- Case, Karl E., and Robert J. Shiller. "Prices of Single-Family Homes Since 1970: New Indexes for Four Cities", *New England Economic Review*, p. 45-56, 1987.
- Dale-Johnson, David, and Hyang K. Yim. "Coastal Development Moratoria and Housing Prices", *Journal of Real Estate Finance and Economics*, Vol. 3, p. 165-184, 1990.
- Delaney, Charles J., and Marc T. Smith. "Impact Fees and the Price of New Housing: An Empirical Study", *AUREA*, Vol. 17, p. 41-54, 1989.
- Dunn, James E. and Lynette Duncan, "Partitioning Mahalanobis D^2 to Sharpen GIS Classification," Management Information Systems Meetings, Lisbon, Portugal, 2000.

- Figlio, David and Maurice Lucas. "What's in a Grade? School Report Cards and Housing Prices", *American Economic Review*, June 2004, vol. 94, p.591-604.
- Gatzlaff, Dean H., and Donald Haurin. "Sample Selection Bias and Repeat-Sales Index Estimates", *Journal of Real Estate Finance and Economics*, Vol. 14, p. 33-50, 1997.
- Gatzlaff, Dean H., and David C. Ling. "Measuring Changes in Local House Prices: An Empirical Investigation of Alternative Methodologies", *Journal of Urban Economics*, Vol. 35, p. 221-244, 1994.
- Goodman, Allen C., and Thomas G. Thibodeau. "Housing Market Segmentation and Hedonic Prediction Accuracy", *Journal of Housing Economics*, 2003.
- Haurin, Donald and David Brasington, "School Quality and Real House Prices: Intra- and Interjurisdictional Effects," *Journal of Housing Economics*, Vol. 5, No. 4, 1996, 351-368.
- Klecka, William R., *Discriminant Analysis*, Beverly Hills: Sage Publications, 1980.
- Leamer, Edward, "Let's Take the Con out of Econometrics," *American Economic Review*, 73, p. 31-43, 1983.
- Leven, Charles L., James T. Little, Hugh O. Nourse, and R. B. Read, *Neighborhood Change: Lesson in the Dynamic of Urban Decay*, New York: Praeger, 1976.
- Lynch. Allen K. and David W. Rasmussen, "Proximity, Neighborhood and the Efficacy Exclusion", *Urban Studies* Vol. 41, February, pages 285-298, 2004.
- Mahalanobis, Prasanta, "On the Generalized Distance in Statistics," *Proceedings of National Institute of Science (India)*, Vol. 12, p. 49-55, 1936.
- Meese, Richard A., and Nancy E. Wallace. "The Construction of Residential Housing Price Indices: A Comparison of Repeat-Sales, Hedonic-Regression, and Hybrid Approaches", *Journal of Real Estate Finance and Economics*, Vol. 14, p. 51-73, 1997.
- Rocke, D. and D. Woodruff, "Identification of outliers in Multivariate Data," *Journal of American Statistical Association*, Vol. 91, No. 435, 1047-1061, 1996.
- Schnare, Ann B. and Raymond J. Struyk. "Segmentation in Urban Housing Markets", *Journal of Urban Economics* 3, 146-166, 1976.

Song, Yan and Gerrit-Jan Knaap, "New Urbanism and Housing Values: A Disaggregate Assessment," *Journal of Urban Economics*, 54, 218-238, 2003.

Straszheim, M.R. *An Econometric Analysis of the Urban Housing Market*, Urban and Regional Studies No. 2, National Bureau of Economic Research: New York, 1975.

Table 1. Results for Full Sample Hedonic Regression

VARIABLE	OLS	SPATIAL LAG	MARGINAL EFFECT
CONSTANT	9.486 (149.67)	9.431 (147.47)	NA
<u>Dwelling Characteristics</u>			
BATHROOMS	0.192 (27.38)	0.192 (27.47)	\$14,798
BEDROOMS	0.045 (7.62)	0.045 (7.56)	\$3,448
CENAIR	0.129 (12.54)	0.129 (12.56)	\$9,911
CENHEAT	-0.017 (-0.92)	-0.018 (-1)	-\$1,381
FIREPLACE	0.101 (14.85)	0.100 (14.74)	\$7,690
LAGE	-0.049 (-10.65)	-0.049 (-10.54)	NA
LOTSIZE	0.098 (12.65)	0.112 (13.79)	\$8,605
PARKING	0.025 (3.05)	0.024 (2.86)	\$1,813
POOL	0.101 (11.12)	0.102 (11.2)	\$7,843
SQFT	0.032 (39.04)	0.032 (38.94)	\$2,430
<u>Neighborhood Characteristics</u>			
AVERAGE INCOME	0.006 (13.99)	0.006 (13.9)	\$494
BLACK%	-0.003 (-14.75)	-0.003 (-14.24)	-\$247
HISPANIC%	0.005 (1.74)	0.005 (1.66)	\$355
OVER50%	0.010 (15.29)	0.010 (15.46)	\$740
OWNER%	0.005 (3.2)	0.005 (3.17)	\$361
OWNER ² %	-6e-005 (-5.21)	-0.0001 (-5.19)	-\$5
POP DENSITY	-0.017 (-4.87)	-0.019 (-5.28)	-\$1,423
WHITE-COLLAR%	0.008 (10.05)	0.007 (9.86)	\$561
<u>Location Characteristics</u>			
DIST_ATLANTIC	-0.0002 (-9.96)	-0.0002 (-9.62)	-\$13
DIST_CBD	0.0002 (4.15)	0.0002 (4.19)	\$18
DIST_STJOHNS	0.0005 (5.37)	0.0005 (5.59)	\$37
WATER FRONT	0.190 (15.69)	0.192 (15.86)	\$14,805
<u>Statistics</u>			
Lag ln(p)	NA	0.006 (5.62)	NA
R ²	0.797	0.798	NA

Notes: The sample size is 7,645. Moran's I-statistic is 0.32, rejecting the null hypothesis of no spatial autocorrelation. The spatial autocorrelation detected has been adjusted using a spatial lag model. T-statistics are reported in parentheses.

Table 2. Regression results using sample segmented by price

VARIABLES	0-10%	10-90%	90-100%
CONSTANT	10.088 (41.52)	10.111 (207.79)	11.539 (40.54)
<u>Dwelling Characteristics</u>			
BATHROOMS	0.019 (0.58)	0.105 (18.72)	0.174 (11.79)
BEDROOMS	-0.001 (-0.01)	0.031 (6.8)	0.011 (0.71)
CENAIR	0.040 (1.74)	0.035 (4.22)	-0.081 (-0.87)
CENHEAT	-0.031 (-0.89)	0.017 (1.17)	0.052 (0.27)
FIREPLACE	-0.026 (-0.85)	0.099 (21.11)	0.026 (0.72)
LAGE	-0.093 (-4.53)	-0.048 (-14.29)	-0.030 (-2.23)
LOT SIZE	-0.111 (-1.81)	0.071 (12.9)	0.082 (4.2)
PARKING	0.091 (1.89)	0.032 (5.67)	-0.067 (-2.16)
POOL	0.189 (1.42)	0.083 (12.44)	0.024 (1.31)
SQFT	0.037 (6.38)	0.028 (42.9)	0.009 (6.14)
<u>Neighborhood Characteristics</u>			
AVERAGE INCOME	-0.002 (-0.7)	0.005 (13.16)	0.003 (2.98)
BLACK%	-0.005 (-7.27)	-0.001 (-4.04)	-0.004 (-2.25)
HISPANIC%	0.001 (0.04)	-0.015 (-7.48)	0.020 (2.37)
OVER50%	-0.001 (-0.06)	0.005 (11.98)	0.007 (3.27)
OWNER%	0.001 (0.24)	0.001 (0.42)	-0.006 (-1.14)
OWNER ² %	0.001 (0.32)	-0.001 (-2.74)	0.001 (0.6)
POP DENSITY	-0.001 (-0.07)	-0.011 (-4.46)	0.010 (0.53)
WHITE-COLLAR%	0.010 (3.66)	0.004 (7.68)	0.001 (0.55)
<u>Location Characteristics</u>			
DIST_ATLANTIC	-0.006 (-1.88)	-0.004 (-12.99)	-0.005 (-2.12)
DIST_CBD	-0.019 (-2.24)	0.005 (4.66)	-0.004 (-0.8)
DIST_STJOHNS	0.034 (2.9)	0.009 (5.08)	-0.004 (-0.48)
WATER FRONT	-0.117 (-0.69)	0.088 (9.16)	0.177 (8.39)
<u>Statistics</u>			
Lagged ln(p)	0.014 (6.12)	0.006 (9.44)	0.003 (1.79)
R ²	0.312	0.755	0.454
N	780	6,101	764

Note: Spatial lag model results are displayed. T-statistics are reported in parentheses. Moran's I for 0-10 data, 10-90 data, 90-100 data is 0.12, 0.27, and 0.44 respectively.

Table 3: Mean Absolute Percentage Error (MAPE) Statistics for Different Mahalanobis selection levels (percentages)

Mahalanobis Cutoff	Covariance											
	0.00			0.25			0.50			0.75		
	α	β	γ	α	β	γ	α	β	γ	α	β	γ
None	4.59	35.12	12.73	7.96	49.44	18.89	12.27	62.96	26.13	17.91	73.65	35.47
		(5.02)			(8.33)			(12.50)			(18.07)	
4.00	1.74	15.77	5.39	3.88	29.55	10.42	6.92	44.30	16.61	11.16	59.95	24.64
		(13.52)			(13.53)			(13.54)			(13.53)	
3.00	0.72	7.08	2.39	2.18	18.39	6.35	4.50	32.72	11.78	7.99	49.44	19.21
		(22.32)			(22.31)			(22.29)			(22.30)	
2.00	0.01	0.20	0.06	0.35	3.62	1.22	1.45	12.99	4.43	3.48	27.04	9.72
		(36.79)			(36.78)			(36.80)			(36.71)	
1.00	0.01	0.31	0.10	0.01	0.32	0.11	0.02	0.35	0.12	0.01	0.47	0.16
		(60.65)			(60.65)			(60.16)			(60.66)	

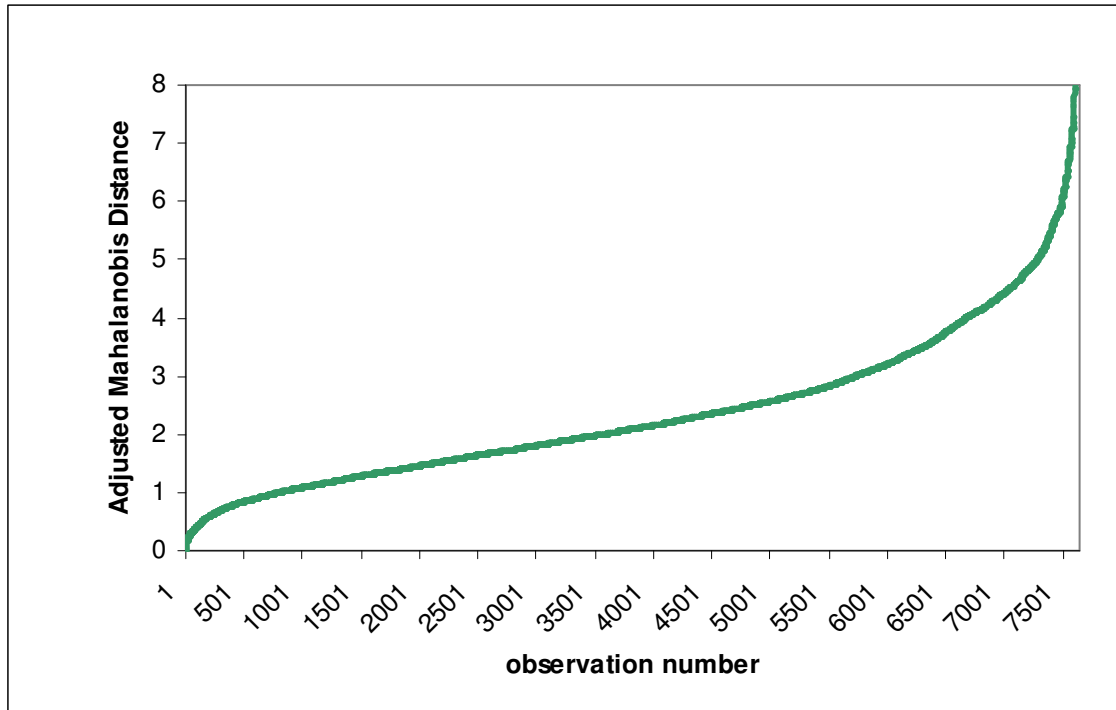
Note: All values reported are percentage deviations from the true coefficients, except values in parentheses that represent the percentage of actual outliers in the first case and the percent of observations cut from the data, when the Mahalanobis sorting is used.

Table 4. Regression results using samples segmented by Mahalanobis method

VARIABLES	Cutoff (Mahalanobis score)			
	1.00	2.00	3.00	None
CONSTANT	10.466 (34.72)	9.582 (77.88)	9.610 (106.88)	9.431 (147.47)
<u>Dwelling Characteristics</u>				
BATHROOMS	0.110 (4.32)	0.144 (13.82)	0.161 (19.40)	0.192 (27.47)
BEDROOMS	0.094 (3.33)	0.060 (7.04)	0.060 (8.78)	0.045 (7.56)
CENAIR	NA NA	0.146 (8.61)	0.117 (10.67)	0.129 (12.56)
FIREPLACE	0.041 (2.46)	0.094 (12.14)	0.099 (15.08)	0.100 (14.74)
LAGE	-0.106 (-8.24)	-0.062 (-10.24)	-0.067 (-13.49)	-0.049 (-10.54)
LOTSIZE	0.246 (4.87)	0.169 (6.13)	0.218 (10.12)	0.112 (13.79)
PARKING	0.029 (1.22)	0.012 (1.15)	0.017 (2.02)	0.024 (2.86)
POOL	NA NA	0.065 (3.83)	0.076 (7.62)	0.102 (11.2)
SQFT	0.040 (18.89)	0.035 (29.73)	0.032 (33.25)	0.032 (38.94)
<u>Neighborhood Characteristics</u>				
AVERAGE INCOME	0.002 (0.79)	0.006 (7.24)	0.008 (12.92)	0.006 (13.9)
BLACK %	-0.003 (-2.33)	-0.004 (-7.66)	-0.004 (-10.41)	-0.003 (-14.24)
HISPANIC %	-0.026 (-3.42)	-0.019 (-4.95)	-0.015 (-4.79)	0.005 (1.66)
OVER50 %	0.005 (2.79)	0.006 (6.29)	0.007 (10.34)	0.010 (15.46)
OWNER %	0.001 (0.03)	0.005 (1.92)	0.006 (2.50)	0.005 (3.17)
OWNER ² %	-0.0001 (-0.42)	-0.0001 (-3.04)	-0.0001 (-4.12)	-0.0001 (-5.19)
POP DENSITY	-0.012 (-1.18)	-0.017 (-3.72)	-0.013 (-3.43)	-0.019 (-5.28)
WHITE-COLLAR%	0.001 (0.39)	0.008 (6.51)	0.006 (6.61)	0.007 (9.86)
<u>Location Characteristics</u>				
DIST_ATLA	-0.0001 (-2.26)	-0.001 (-4.75)	-0.0001 (-7.76)	-0.0002 (-9.62)
DIST_CBD	-0.0001 (-0.67)	-0.001 (-1.82)	-0.0001 (-1.85)	0.0002 (4.19)
DIST_STJO	-0.0001 (-0.69)	0.004 (5.35)	0.0008 (7.26)	0.0005 (5.59)
WATER FRONT	NA NA	NA NA	0.027 (0.72)	0.192 (15.86)
<u>Statistics</u>				
Lag ln(p)	-0.002 (-1.56)	0.002 (2.44)	0.002 (1.54)	0.006 (5.62)
R ²	0.7878	0.7876	0.7979	0.798
N	765	3,581	5,352	7,645

Note: T-statistics are reported in parentheses. NA indicates variables with no variation. Moran's I-statistic is 0.012, 0.261, 0.004, and 0.320 for the above samples. Column 4 is identical to the second column in Table 1.

Figure 1. Houses Sorted by Mahalanobis Distance



Note: The adjusted Mahalanobis distance is the difference between the Mahalanobis distance for a given house and the minimum Mahalanobis distance for the entire data set. To make the graph easier to see, thirty-nine observations have been deleted from the above graph, because their adjusted Mahalanobis distance is too large exceeding 8 and ranging up to 31.

Appendix A. Definition and summary statistics of variables

VARIABLE	DEFINITION	MEAN	MINIMUM	MAXIMUM
AVERAGE INCOME	Average income within a half-mile radius	40.47	12.55	105.16
BATHROOMS	Number of bathrooms	1.87	1.00	7.00
BEDROOMS	Number of bedrooms	3.06	1.00	6.00
BLACK%	Percentage of African-Americans within a half-mile radius	15.10	0.00	100.05
CENAIR	1 if the dwelling has central air-conditioning, 0 otherwise	0.87	0.00	1.00
CENHEAT	1 if the dwelling has central heating, 0 otherwise	0.97	0.00	1.00
DIST_ATLANTIC	Distance to Atlantic (in miles)	13.80	0.01	37.81
DIST_CBD	Distance to Central Business District (in miles)	8.46	0.02	23.85
DIST_STJOHNS	Distance to St Johns River (in miles)	2.85	0.00	19.02
FIREPLACE	1 if the dwelling has one ore more fireplaces, 0 otherwise	0.57	0.00	1.00
HISPANIC%	Percentage of Hispanics within a half-mile radius	2.62	0.00	8.11
LAGE	Natural log of the age of the dwelling	2.90	0.69	4.59
LOTSIZE	Number of acres	0.31	0.04	11.73
LPRICE	Natural log of the selling price	11.25	8.88	13.74
OVER50%	Percentage of people who are 50 and over within a half-mile radius	21.32	3.13	48.08
OWNER%	Percentage of owner-occupied units within a half-mile radius	72.42	12.87	97.72
OWNER ² %	Squared term of the percentage of owner-occupied units within a half-mile radius	5442.20	165.74	9549.80
PARKING	1 if the dwelling has a garage, 0 otherwise	0.62	0.00	1.00
POOL	1 if the dwelling has a pool, 0 otherwise	0.12	0.00	1.00
POP DENSITY	Population density within a half-mile radius	1.91	0.04	7.38
SQFT	Base square footage of living space (in hundreds)	14.28	2.64	49.10
WATERFRONT	1 if the dwelling is waterfront property, 0 otherwise	0.06	0.00	1.00
WHITE-COLLAR%	Percentage of white-collar workers within a half-mile radius	39.53	13.54	57.88

Appendix B. Summary statistics for Sub-samples

VARIABLES	10% (n=765)			46% (n=3,581)			70% (n=5,352)		
	Mean	Minimum	Maximum	Mean	Minimum	Maximum	Mean	Minimum	Maximum
AVERAGE INCOME	42.29	22.65	61.43	40.62	20.34	72.51	40.46	17.81	78.28
BATHROOMS	1.91	1.00	2.50	1.83	1.00	3.50	1.82	1.00	3.50
BEDROOMS	2.99	2.00	4.00	3.02	2.00	4.00	3.03	2.00	5.00
BLACK%	8.93	2.07	43.42	11.47	0.00	65.21	12.54	0.00	83.95
CENAIR	1.00	1.00	1.00	0.96	0.00	1.00	0.92	0.00	1.00
CENHEAT	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
DIST_ATLANTIC	12.66	2.73	24.09	14.26	0.37	26.45	13.89	0.09	27.01
DIST_CBD	9.36	2.48	14.54	8.49	2.02	15.34	8.44	1.00	16.18
DIST_STJOHNS	2.43	0.14	6.20	2.62	0.00	7.82	2.65	0.00	8.38
FIREPLACE	0.76	0.00	1.00	0.60	0.00	1.00	0.58	0.00	1.00
HISPANIC%	2.92	0.86	5.29	2.85	0.37	6.40	2.74	0.00	6.40
LAGE	2.62	1.10	4.33	2.80	0.69	4.45	2.87	0.69	4.49
LOTSIZE	0.24	0.08	0.73	0.25	0.05	1.29	0.25	0.05	1.55
LPRICE	11.27	9.85	12.00	11.21	9.21	12.91	11.22	9.11	12.91
OVER50%	17.76	9.55	31.68	19.71	9.33	38.79	20.54	9.33	41.62
OWNER%	74.93	48.33	92.13	74.28	40.01	94.49	73.60	37.51	97.67
OWNER ² %	5702.00	2335.65	8488.09	5661.68	1600.88	8927.75	5573.20	1406.71	9540.29
PARKING	0.83	0.00	1.00	0.69	0.00	1.00	0.66	0.00	1.00
POOL	0.00	0.00	0.00	0.03	0.00	1.00	0.08	0.00	1.00
POP DENSITY	1.83	0.33	4.35	1.95	0.07	5.69	1.96	0.07	5.74
SQFT	14.52	6.72	21.45	14.07	5.00	28.83	14.08	4.30	30.09
WATERFRONT	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	1.00
WHITE-COLLAR%	41.39	31.91	47.64	40.29	24.55	49.93	40.07	21.09	52.40

Note that the sub-samples refer to the samples used in Table 3.