THE INFLUENCE OF WATER QUALITY ON THE DEMAND FOR RESIDENTIAL DEVELOPMENT AROUND LAKE ERIE

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the Graduate

School of The Ohio State University

By

Shihomi Ara, B.A., M.A.

*****

The Ohio State University
2007

Approved by

Dissertation Committee:
Professor Timothy Haab, Advisor _____

Professor Elena Irwin, Co-Advisor _____

Professor Brent Sohngen

Graduate Program in
Agricultural, Environmental
and Development Economics

ABSTRACT

In early 1970s, it was said that Lake Erie was dead. In 1950s and 1970s, the water of the Lake was pea-green colored due to excessive phosphorous from sewage and runoffs from farmlands and homeowners. There were many closed beaches and fish from the Lake was not edible. However, water quality has improved dramatically since the Clean Water Act of 1972. The pace of residential and commercial development around the shoreline of Lake Erie increased considerably following substantial improvements in the Lake's water quality and clarity. A double-edged sword exists since increases in water quality are followed by increases in residential and lake-related development, which in turn can degrade the lake and the amenities it provides. In fact, although the phosphorous level declined in 1980s, we are observing an increasing trend starting from 1990s and the trend continues until today. As for water clarity, although its level hit the peak in 1995, we observe the decreasing trend afterwards. In this study, we focus on the effects of water quality on housing values to evaluate water quality-housing value as the relationship on the one side of the double-edged sword.

Both the first and the second stage of hedonic price analysis are conducted with identified housing submarkets by using Hierarchical Clustering with quantized similarity measures in the region including Erie, Lorain, Ottawa and Sandusky Counties located

along Lake Erie. We use both individual houses and census block groups as the smallest building blocks of the clusters and compare the clustering and hedonic results for both cases.

Fecal coliform counts and secchi disk depth readings measuring water clarity are used as water quality variables. In order to overcome the spatio-temporal aspects of secchi depth disk reading data, kriging was used for spatial prediction. Robust Lagrange Multiplier test indicates that spatial error models are appropriate for the estimation of hedonic price functions in each submarket. We found that secchi disk depth readings variables are positive significantly influencing housing prices in most of the clusters while mixed results are found for fecal coliform counts.

Marginal implicit prices (MIP) are computed based on the estimated results of the first stage hedonic price functions. As for the houses whose prices are negatively influenced by fecal, the MIP for reducing the amount of bacterial counts is estimated as -21.6 dollars (in 1996$), and -30.5 dollars for the houses affected by fecal statistically significantly. For water clarity, MIP is estimated as 40.5 dollars for the houses whose housing price is positively affected by the variable, and it is 56 dollars for the houses significantly affected by water clarity.

Demand functions with different functional forms are estimated with two-stage least squares with submarket dummy variables. We found that fecal coliform and water clarity are substitutes to each other while the distance to the closest beach is a complement to fecal coliform and substitute to water clarity. While computed welfare changes for fecal coliform by using non-linear demand functions are very small, the benefit of the

improvement of water clarity by 25 centimeters to be estimated 230 dollars per household. We found that the welfare changes are larger for the degradation of water quality compared to the improvements of water quality in the same amount.

We further analyzed the welfare changes by using demand functions derived specifically for each household. Welfare changes based on the individual demand functions were computed by integrating under each demand curve for multiple scenarios. If we consider our SIG Fecal data represents 33 percent of entire population in four counties, the total estimated net benefit was derived as 51,934,180 dollars for targeting 155 fecal coliform counts. The total net welfare gain was computed as 899,010,835 dollars for targeting 245 centimeters of water clarity.

To my husband, Selim,
parents, Shigeo and Emiko and sister, Masumi

ACKNOWLEDGMENTS

I wish to express my deep gratitude to Professor Elena Irwin for her constant encouragement and patience throughout the entire process of my dissertation work. Without her warm support, I could not complete my dissertation. I would like to thank Professor Timothy Haab for his support from very beginning of my graduate study. I also wish to thank my committee member, Professor Brent Sohngen for his suggestions and comments for the prospectus and this dissertation.

My big appreciation also goes to Ms. Jill Clark for her assistance regarding ArcGIS. Her patience in teaching me operations in Arc especially in the early stage of the project and her prompt and helpful responses to my requests are highly appreciated. I also would like to thank Professor David Culver, Department of Evolution, Ecology and Organismal Biology, the Ohio State University for providing useful advice regarding to the water quality data.

I also would like thank Ms. Susan Miller for her help in many ways from the very beginning of my graduate work. Without her help, especially when I was out of the U.S, I could not complete necessary administrative works in time.

Through the course of my graduate work, I have met many fellow graduate students. I would like to thank my classmates for going through Ph.D works together with nice cooperation and kind supports. I am also indebted to Nobuyuki Iwai, Naoya Kaneko, Takeshi Miyata, Shigeharu Okajima and other Japanese friends for offering me cordial friendship.

The friendships outside AEDE also helped me great deal in pursuing my dreams. Without their encouragements and kindness in various ways, I could not sustain the enjoyable aspects of my Ph.D student life. Thank you, Zehra, Zeynep, Gulcin, Nilgun, Nilufer, Sefa Nur and Zehra for your beautiful friendships.

I would like to thank my parents, Shigeo and Emiko Ara, and sister, Masumi Ara for giving me their ungrudging support throughout my academic study. Last, but never the least, I would like to thank my husband, Dr. Selim Aksoy for encouraging me to pursue my goals both in good days and bad days. Without his dedicated support, it was impossible to complete my dissertation.

This research was supported by Ohio Sea Grant. I am grateful for the financial support. I am also thankful for the travel funding support from the Environmental Policy Initiatives of OSU for making my presentation in 3rd World Congress of Environmental Economics in Kyoto possible. I also appreciate AAEA for providing the travel fund for my presentation.

I would like thank the following organizations for providing us data necessary to conduct our study: CURA (Center for Urban and Regional Analysis), the Ohio State University for housing data, Ohio Department of Health for Fecal coliform counts data, and Stone Laboratory of the Ohio State University and Ohio Department of Natural Resources for Secchi disk depth readings data.

VITA

July 4, 1976. . . . . Born, Fukuoka, Japan

2000. . . . . B.A., AoyamaGakuin University, Japan

2002. . . . . M.A., Kobe University, Japan

2003. . . . . M.A., The Ohio State University

2001 – 2006. . . . . Graduate Research Assistant.
Department of Agricultural, Environmental and
Development Economics The Ohio State University.

PUBLICATIONS

FIELDS OF STUDY

Major Field: Agricultural, Environmental and Development Economics
Studies in: Environmental Economics, Nonmarket Valuation

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Lake Erie is one of the five large freshwater lakes in North America and the 13th largest natural lake in the world. In the early 1970s, it was said Lake Erie was dead. The Lake was experiencing algae boom due to excessive phosphorous from sewage and runoffs from farmlands and homeowners. In 1950s and 1970s, the water of the Lake was pea-green colored. There were many closed beaches and fish from the Lake was not eatable. In 1969, Cuyahoga River which flows into Lake Eire through Cleveland caught on fire due to its polluted water covered with oil. This event called policy makers' attention and led to the Great Lake Water Quality Act and Clean Water Act in the 1970s.

The sources of pollution back then were point source pollution such as industrial water discharge and emission from sewage treatment plants as well as non-point source pollution such as runoff from upstream agricultural lands containing fertilizers and pesticides.

Since the Clean Water Act of 1972, point source pollution has been controlled well. Nearly 100 percent of all industrial plants use control measures to reduce their toxic discharge, and the number of sewage treatment plants has doubled. Water quality of Lake Erie has improved since then. In 1983, phosphorous level in the Lake met its standard.

The pace of residential and commercial development around the shoreline of Lake Erie increased considerably following substantial improvements in the lake's water quality and clarity in the 1970's and 1980's. Between 1982 and 1997, the amount of urban land use in the eight Ohio counties bordering Lake Erie increased 24.4%, an increase of 112,500 acres (USDA, National Resources Inventory). A significant portion of this development appears tied to Lake Erie. For example, the amount of urban development in Ottawa County, a county that contains a number of lake amenities and recreational sites, increased 53% during this fifteen year time period.

The improvement in water quality enhanced the urban development along lakeshore counties. On the other hand, the impacts of urbanization and development in coastal areas threaten the very resources that make these areas attractive as places to live. The cumulative effects of non-point source pollution on coastal waters and aquatic life is a critical and increasing concern both nationally and in the Lake Erie watershed. Like many streams and rivers in the U.S., sedimentation and hydro-modification are cited as primary non-point sources affecting Lake Erie tributaries (Ohio EPA, 1996). Of particular concern is urban storm water pollution, which is fast becoming the most serious type of water pollution affecting Ohio's streams and near shore areas. This

source of non-point pollution has been driven by the conversion of farmland and forests to urban uses. These trends point to a complex relationship between lake quality and surrounding urban land development. Increases in lake quality are followed by increases in lake-related development, which in turn can degrade the lake and the amenities it provides. This double-edged sword presents policymakers with a tough challenge: how to attain improvements in lake quality and manage the increased lake-related development that often follows. In fact, according to the study conducted by the USEPA's Great Lakes National Program Office, phosphorus level began to increase again after the reduction in 1980s and the increasing trend has continued to the present day (USEPA (2006)). Water clarity has also been decreasing up to the present day after marking the peak value in 1995.

In this research we focus on the linkage between lake quality and housing values and how the lake quality influences the demand for residential housing. Through the use of extensive data available on housing transactions in the Lake Erie watershed, we will estimate the effects of changes in water quality on housing prices by using the first stage of hedonic price estimation, and demand for water quality and welfare changes due to the changes in water quality by conducting the second stage of hedonic price analysis.

Cluster Analysis is adopted in order to identify distinguishable submarkets existing in the extent of our housing data. Four similarity measures which incorporate the mixed (continuous and categorical) features of our clustering variables are used in order to implement hierarchical clustering.

Two indicators of water quality are included. One is fecal coliform counts which are a basis for beach closing and posting of warning on each beach, and the other is secchi disk depth readings which are an indicator of water clarity of the lake water. Kriging is done in order to overcome the handling of secchi disk data which are collected over different points in time and space.

Given ten identified clusters, we estimate spatial hedonic functions for each cluster. Multiple robust Language Multiplier (LM) tests are conducted and appropriate spatial models and weight matrices specifications are chosen.

The estimated results are used to compute the marginal willingness to pay (WTP) or implicit prices of water quality in the area for different sets of houses whose sales prices are affected by water quality. In other words, we know how much individual house owners are willing to pay for one unit increase in water clarity, or one unit decrease in bacterial counts.

The procedure for estimating demand function in order to compute the welfare change for a non-marginal change in water quality will be implemented in the second stage of the hedonic method. Demand identification problem is dealt with estimation of multiple hedonic price functions from separate submarkets determined by Cluster Analysis. When hedonic price function is non-linear, households face to choose quantity and price of the water quality at the same time. In order to resolve the endogeneity problem, two-stage least squares estimation method with instrumental variables (IVs) will be employed.

Identified ten clusters are considered to form separate housing submarkets and included into the second stage hedonic analysis in this research and demand functions are identified. Welfare changes due to the non-marginal changes in water quality are computed as the last stage of this hedonic study. Furthermore, we derived the individual demand functions and the individual welfare changes for multiple scenarios of water quality changes. The welfare changes are calculated by using the actual water quality values each household is experiencing. The total net welfare gains for the relevant population are reported in the end.

Hedonic price analysis reveals the value of environmental amenity (water quality in our case) through the preference of house seekers/owners and transactions in housing market. Therefore, the benefits and welfare measures derived in the end of the hedonic analysis do not reflect all aspects of the value of environmental amenity, but capture simply some portions that are perceived by home purchasers. Therefore, we have to emphasize that the benefits we estimate are based on the house purchasers' perception regarding the Lake water quality, and do not include the benefits possibly obtained in other ways.

Furthermore, the benefits are measured from a human point of view, not lake biological point of view. In other words, what is desired by human may not coincide with the ideal environment for the Lake and the creatures living there. In order to capture the entire service an environmental amenity provides, we should conduct other types of research, such as Travel Cost Method as the other revealed preference method and/or Contingent Valuation Method or Conjoint Analysis as stated preference methods.

Therefore, it is important to keep in mind when we interpret the results of this research that the welfare measures we estimated in the end of the second stage hedonic analysis are partial benefits we obtain from Lake Erie water and its quality.

The main contributions of our study include 1) an introduction of similarity measures which handle mixed featured variables more precisely than widely used Euclidean distance and implementation of hierarchical clustering with suggested similarity measures, 2) comparison of submarket definition by using two different building blocks, individual house and census block groups, 3) spatial data handling of water clarity data by using kriging method, 4) estimations of marginal WTP for water quality on each beach along Lake Erie by using spatial error models, 5) derivation of demand functions for water quality and calculation of welfare changes due to non-marginal changes in water quality and 6) derivation of individual demand function and welfare changes based on individual specific variables and computation of total net benefits of water quality for relevant population.

CHAPTER 2

OVERVIEW OF THE HEDONIC METHOD

2.1 Related Works on Hedonic Method

Many hedonic price studies have conducted in the past. We introduce major studies using water quality as well as other environmental variables in this section. The number of hedonic studies involving water quality itself is small comparing to air-quality studies. Hedonic studies with water quality considering spatial effects are very limited.

2.1.1 Applications of Hedonic Method with Environmental Variables

Boyle and Kiel conducted a survey on house price hedonic studies considering the impact of environmental externalities (Boyle and Kiel (2001)). According to their survey, hedonic pricing model has been used to evaluate the impacts of environmental goods such as air quality, water quality, undesirable land use and multiple environmental goods.

As we already mentioned earlier, the earliest hedonic air quality study was done by Ridker and Henning (1967). Sulfation levels were used in the final report. They report that a decrease in sulfation of 0.25 mg/ 100 cm^2/day would increase values of owner-occupied single family houses by $83 - $245 in 1960 dollars. Wieand (1973) used suspended particulates, sulfur dioxide and sulfur trioxide as air quality measures and

estimated their effects on monthly rent per acre of land by using Ridker and Henning's data. He showed that there is no statistically significant effect of air pollution on monthly rent. Smith and Deyak (1975) used suspended particulates as air quality measure and looked at eighty five central cities in U.S. They did not find statistically significant effect. Harrison and Rubinfeld (1978) included squared NO_2 concentration from a meteorological model. The estimated results for NO_2 concentration were negative and statistically significant. Nelson (1978) used particulate concentration and summer oxidant concentration as pollution measures. The coefficient for particulate concentration was negative and statistically significant. Implicit prices calculated were $57.61 for particulates and - $14.11 for oxidants in 1970 dollars. Li and brown (1980) included sulfur dioxides and total suspended particulates as environmental variables and obtained marginally statistically significant result for both variables in one of the models estimated. Palmquist (1982) included total suspended particulates (TSP), nitrogen dioxide, ozone and sulfur as air quality measures, used house sales prices as the dependent variable and obtained mixed sign and statistical significant results. Palmquist (1983) included pollution measures as an index instead of four different measures as in his 1982 study and got statistically significant results in six of the fourteen cities. Murdoch and Thayer (1988) used mean visibility and sales data and obtained positive statistically significant result. Graves *et al* (1988) included particulates and an index of visibility and reported negative and significant results for particulates and greatly various results for the index of visibility. They also mentioned that hedonic results are sensitive to the choices of right hand side variables. Zabel and Kiel (2000) used the arithmetic

mean of nitrogen dioxide readings and sulfur dioxide readings, and the second daily maximum hourly readings for ozone and total suspended particulates for four urban areas in the U.S. over five time periods. They obtained mixed results for pollution coefficients.

Anselin (2004) studied effect of air quality on house price in L.A. in 1999 by using different spatial interpolation methods (Thiessen Polygon, Inverse Distance Weight and Kriging) for O3 readings from 27 monitoring stations. They concluded that different spatial interpolation methods give different estimated results and kriging is the best measure to include.

Blomquist (1974) included the effective distance to an electrical power plant as the environmental variable. Effective distance equals to distance if the distance was less than 11,500 feet and equal to 11,500 if the distance is greater than the value. If the property was 10% further away from the plant, an average value of the property increased by 0.9% within 11,500 feet of the plant. Nelson (1981) studied the impact of the nuclear power plant accident at Three Mile Island on house prices. He did not obtain statistically significant results on the value of interest. Gamble and Downing (1982) also conducted a study on nuclear reactors. They looked at two cases, one plant without accident and another (Three Mile Island) with accident. For the first case, two variables for the environmental measure were included. One dummy variable which equals to one if the plant was visible from the house and a variable of distance measure from the plant. They did not find statistically significance for either variable. For the second case, distance from the house to the plant, dummy variable for whether the house was purchased after the accident, and the interaction of the two variables were included. Distance from the

plant was significant before the accident, but not after. McClelland, Shulze and Hurd (1990) looked at the perceived risk of living near a hazardous waste site. Their study is based on mail survey and authors developed an estimate for perceived risk. An increase of 10% in the proportion of survey participants rating an area as "high risk" would decrease on average nominal sales price of about $2,084. Closing the landfill increased average house value by about $5,001. After closing the land fill, average house prices were estimated $4,793 lower than the case with no perceived health risk.

Other undesired land uses employed in hedonic studies are Superfund site (Kohlhase (1991), Keil (1995)), landfill (Nelson, Genereux and Genereux (1992), Reichert, Small and Mohanty (1992), Smolen, Moore and Conway (1992)), hazardous waste sites (Michaels and Smith (1990), Ketkar (1992)), petroleum refineries (Flower and Ragas (1994)), incinator (Keil and McClain (1995)), lead smelter (Dale *et al*. (1999)), and petroleum pipeline rupture (Simons (1999)).

2.1.2 Applications of Hedonic Method with Water Quality Variables

David (1968) is the first study of hedonic water quality analysis. She designated the water quality in each lake as poor, moderate or good based on the specialists' opinions for sixty artificial Wisconsin lakes. Dependent variables used are per acre value of land, per acre value of improvements, per acre number of improvements, and land value calculated as a weighted sum that relates per acre land value to per acre value of improvements and per acre number of improvements. She concluded that property on more polluted lakes was less valuable than property adjacent to cleaner lakes.

Epp and Al-Ani (1979) incorporated water pH and interaction between pH and percentage change in population for the years 1960-1970 as the water quality variables as well as perceived water quality and the interaction of perceived water quality with percentage change in population. Property values in Pennsylvania from 1969 to 1976 were dependent variable. Independent variables include flood hazard, lot size, the number of rooms, potential employment, school expenditure per pupil and age of the house. They concluded that both measures of water quality have statistically significant effect on property values. A one-point increase in pH would result in $653.96 (1972 $s) increase in the mean sales value of the properties. The interaction variable was significant only for good quality steam case. They also found that for the case of poor quality streams, property characteristics other than water quality variables have greater importance to purchaser of the property.

Young (1984) included one to ten water quality ratings by local officials and dummy variable indicating if the house is adjacent to St. Albans Bay where a malfunctioning waster treatment plant had caused pollution problems. The estimation result shows that a location within the bay reduced property values by an average of 20%. If the properties are adjacent to the bay, the value was an average of $4,700 less than equivalent properties.

Steinnes (1992) studied fifty-three Minnesota lakes by using secchi depth disk readings as his water quality measure. The author used three dependent variables, total price of all lots on the lake, average price per lot on the lake and average price per front foot of lot on the lake. He found that each additional foot of clarity would raise the value of a lot by $206.

Mendelsohn *et al.* (1992) looked at PCB pollution in the New Bedford, Massachusetts harbor. They employed sales data from 1969 to 1988 and used panel data approach. They included a dummy variable to indicate if the sale occurred before or after the PCB pollution and interactions between the year and two dummy variables to indicate homes whose nearest waters were affected by PCB pollution. The properties affected by PCB pollution had $7,000 to $10,000 (1989 $s) lower values.

Michael, Boyle and Bouchard (1996) used secchi depth disk readings of minimum clarity for thirty-four Maine lakes. They used property sales records between January 1st, 1990 and June 1st, 1994. They found that a one-meter improvement in lake clarity would increase property prices by anywhere from $11 to $200 per foot frontage.

These studies have generally demonstrated a positive relationship between water quality/lake amenities and residential property values. But none of them paid enough attention to the spatial feature of water quality data. As the study by Anselin (2004) indicates, the estimated result could change depending on how researchers treat and include environmental variables in their estimation. Therefore, we would like to take closer look at spatial and spatio-temporal features of water quality data and study how much estimated results could be affected by including different type of data. Furthermore,

none of above studies has extended the analysis to consider the implications of these positively valued lake attributes for future development patterns and for lake and land use policies. This research will do both by combining the traditional hedonic model with experiments that will examine various water quality and lake amenity scenarios and that will allow us to extend the analysis to consider these additional questions.

2.1.3 Applications of Spatial Hedonic Method

Kim, Phipps and Anselin (2003) developed spatial-econometric hedonic housing price model to estimate for the Seoul metropolitan area to measure the marginal value of improvements in sulfur dioxide (SO_2) and nitrogen dioxide (NOx) concentration. Their test result favored the spatial-lag model over the spatial error model. The estimated model is

$$\mathbf{P} = \rho\mathbf{WP} + \mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{X_3}\boldsymbol{\beta_3} + \varepsilon$$

where $\mathbf{P}$ is the vector of housing prices, ρ is a spatial autocorrelation parameter, $\mathbf{W}$ is a n by n spatial weight matrix (where n is the number of observations), $\mathbf{X_1}$ is a matrix for structural characteristics, $\mathbf{X_2}$ is a matrix for neighborhood characteristics, and $\mathbf{X_3}$ is a matrix with observations on environmental quality variables, with e assumed to be a vector of independent and identically distributed (i.i.d.) error terms.

The authors used semi-log specification and compared the estimated results from four different estimation methods, OLS, ML, Spatial two-stage-least squares (S-2SLS), and heteroskedastic robust spatial two-stage-least squares (robust S-2SLS). They found that Marginal WTP for a permanent 4% improvement in air quality is about $2,333 for owners by using robust S-2SLS. They also found that OLS overestimates the welfare measure.

Leggett and Bockstael (2000) employed spatial error model with entries of weight matrix being zero if the distance between two houses exceed one mile. Inverse distance-weighted average of fecal coliform counts was used as their water quality measure and controlled for emitter effects by including straight-line distance to the nearest sewage treatment plant to investigate the influence of water quality on residential property values of houses along the Chesapeake Bay coastline. They found that a change of 100 fecal coliform count /100 mL resulted in a change in property prices of about 1.5%.

Beron *et.al*. (2003) implemented spatial error model by incorporating particulate matter of size 10 microns or less (PM 10) as the air quality variable. They also controlled heterogeneity in the model by including the quadratic expansion of the X, Y coordinates to model the spatial trend. The air quality was proved to be significantly affecting the housing price in four counties in Southern California.

2.2 First Stage Hedonic Method

A differentiated good is a good which is composed of multiple characteristics. If the good is a computer, its characteristics include CPU, RAM, OS, memory, and more. When we talk about a house as the good, the characteristics contains housing characteristics such as size of the house, size of the lot, number of bedrooms, number of bathrooms, and size of the garage, as well as neighborhood and proximity characteristics such as school district ranking, average income level of the neighborhood, crime rate, proximity to major cities and so forth. The hedonic price method is a technique to estimate implicit prices of the characteristics of a differentiated good. The partial derivative of the estimated hedonic price function with respect to a characteristic gives the marginal implicit price. The marginal implicit price indicates how much the price of the good is affected by one unit change in the characteristic. We focus on housing markets in the following sections and consider the case where the environmental quality around the house is considered one of the characteristics determining the housing price.

2.2.1 Theory

Let $\mathbf{Z}$ be a vector of housing characteristics, consisting of $z_1, z_2, ..., z_3, E_1, E_2$ where E_k represents environmental variables. The hedonic price function is expressed as $P(\mathbf{Z}) = f(z_1, z_2, ..., z_n, E_1, E_2)$. This is an equilibrium price schedule for the differentiated good derived under the assumption that the good is sold in a perfectly competitive market with the interactions of many consumers and producers. It is important to note that the entire price schedule $P(\mathbf{Z})$ is exogenous to the consumers, but consumers can determine

how much they pay for the good by choosing which good (e.g. house) with certain characteristics to purchase (Taylor 2003). Assume now that an individual purchases only one house in a certain time period. The consumer j's utility function can be written with two parts, a house with various characteristics and the numeraire good, X given her demographic characteristics **C**.

$$U_j = U_j(\mathbf{Z}, X; \mathbf{C}_j) \qquad (2.1)$$

Since we assume that the consumer purchases only one unit of the good, her budget constraint can be expressed as the separable form as follows.

$$Y_j = P(\mathbf{Z}) + X \qquad (2.2)$$

Plugging (2.2) into (2.1) gives

$$U = U(z_1, z_2, ..., z_n, E_1, E_2, Y - P(\mathbf{Z})) \quad . \qquad (2.3)$$

Inverting (2.3) by holding all constant except for the characteristics i gives us a bid curve which provides the maximum amount the individual would pay to obtain the specific house as a function of z_i or E_k, $\mathbf{Z}_{-i}$* and U*, where $\mathbf{Z}_{-i}$* indicates the optimal level of other characteristics chosen and U* represent the maximized level of utility as the solution to utility maximization problem (Freeman 1993). The bid function can be expressed as

$$B_i = B_i(z_i; \mathbf{Z}_{-i}^*, U^*, Y, \mathbf{C}) \, . \qquad (2.4)$$

Since we do not control for individual characteristics such as income and preference in the first stage of hedonic method, we can derive a different bid function for each individual. Maximizing (2.1) subject to the budget constraint (2.2) gives the condition for an individual to choose the levels of each characteristic as follows.

$$\frac{\partial U / \partial z_i}{\partial U / \partial X} = \frac{\partial P(\mathbf{Z})}{\partial z_i} \qquad (2.5)$$

Similarly, firms' offer curves can be derived as follows. A farm maximizes their profits $\Pi = Q * P(\mathbf{Z}) - C(Q, \mathbf{Z}, \mathbf{S})$, where Q is the number of unites of Z the firm produces, C(.) is a cost function, **S** is the firm's characteristics. We assume that each firm has a different cost function. By inverting the profit function at the optimal level, we can derive the offer function as

$$C_i = C_i(z_i; Q^*, \mathbf{Z}_{-i}^*, \Pi^*, \mathbf{S}) \qquad (2.6)$$

where Π^* is the maximum level of profits. A hedonic price function is derived as a double envelope of the sets of bid and offer functions (Rosen 1974). The illustration is shown in Figure 2.1. In Figure 2.1, bid and offer functions for the housing characteristics i for two consumers and firms are shown.

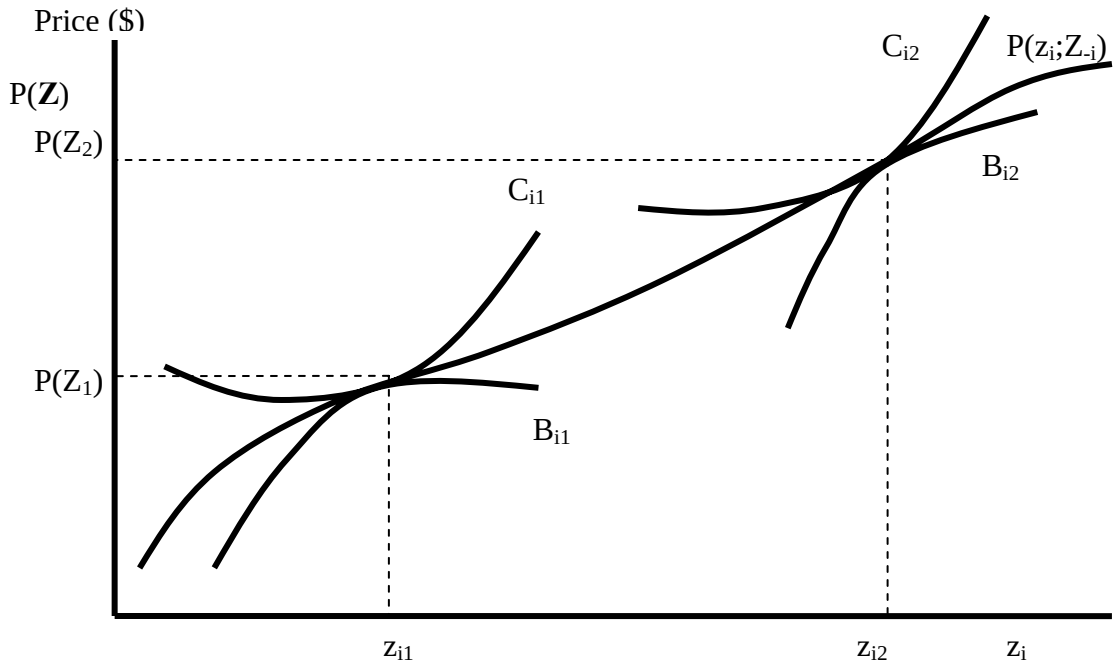


Figure 2.1. Hedonic Price Function for Characteristic i
as the Envelope of Bid and Offer Functions.

Implicit price function can be derived by taking a derivative of the hedonic price function with respect to z_i. For the given level of z_i (e.g. z_{i1} for consumer 1 in Figure 2.1.), we can recover one point on the implicit price function where individual willingness to pay function intersect (See Figure 2.2.). It is the point A in Figure 2.2 for individual 1 and point B for 2. Point A represents the point where the condition in (2.5) is met and is equal to the marginal willingness to pay (WTP) of consumer 1 for the characteristic i. b_{ij} is marginal bid or WTP function. It is also equivalent to an inverse compensated demand function for the characteristic i and shows the change in WTP for z_i for the marginal change in quantity of z_i, holding utility at the maximized level and all

other characteristics constant (Taylor, 2003). Since we do not control for the individual characteristics in the first stage, we obtain multiple points on the implicit price function for each individual. However, due to the lack of information for determining further the shape of the marginal bid function, all we can obtain at this point is not a function b_{ij} as depicted as dotted line, but simply a point on implicit price function such as point A and B. Determination of function b_{ij} is dealt in section 2.4 below.

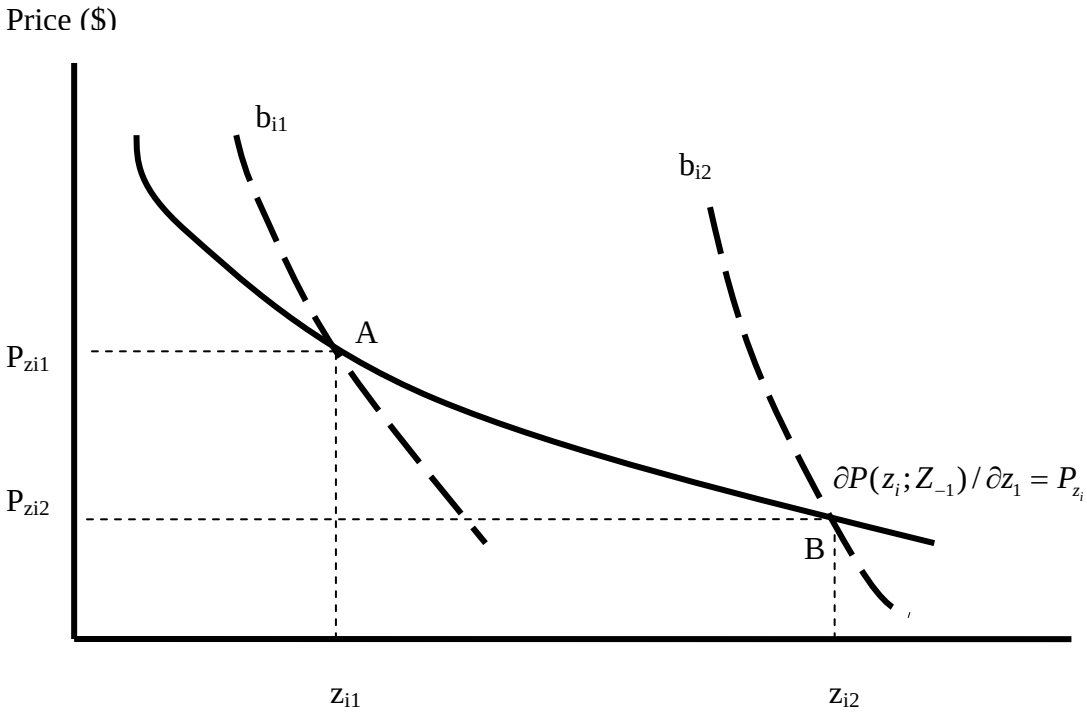


Figure 2.2 Implicit Price Function and Individual Willingness to Pay

2.2.2 The Model

Sales prices, household or tax assessor values, and rental prices of the properties are typically used as the dependent variable of the hedonic price function. Independent variables in hedonic price function are the ones considered to affect housing prices. As we stated in the previous section, consumer and producers' characteristics do not enter in the regression.

General hedonic price models have employed different functional forms to estimate the effects of independent variables (housing structures, neighborhood environments, proximity to places, other variables of interests such as environmental variables and crime rate on property values. General form could be expressed as

$$P = P(\mathbf{H}, \mathbf{N}, \mathbf{D}, \mathbf{E}) \qquad (2.7)$$

where P is the sales price of a house, $\mathbf{H}$ is structural and property characteristics of the house, such as number of bedrooms and lot sizes, $\mathbf{N}$ represents neighborhood characteristics, such as school district ranking, median income in a census block group, racial composition, $\mathbf{D}$ is proximity to places, such as proximity to urban center, big cities and beaches, and $\mathbf{E}$ represents environmental variables or other variables of interests.

The possible functional forms for the hedonic price function are listed in Table 2.1. According to the study on functional form by Cropper, Deck and McConnell (1988), when all attributes of housing are observed without error, the complicated functional forms, such as quadratic, linear Box-Cox, and quadratic box-Cox can be used to estimate implicit prices more accurately. However, when some variables are not observed or are replaced by proxy variables, simpler forms such as linear, semi-log, double-log and linear

Box-Cox are preferred since the quadratic and quadratic Box-Cox produced biased estimates of the marginal prices. They concluded that linear Box-Cox is the most preferable functional form since it provides accurate marginal price estimates when all attributes are measured correctly and also performs well in the presence of mis-specification of the hedonic function.

Name	Equation	Implicit Prices
Linear	$P = \alpha_0 + \sum \beta_i z_i$	$\partial P / \partial z_i = \beta_i$
Semi-Log	$\ln P = \alpha_0 + \sum \beta_i z_i$	$\partial P / \partial z_i = \beta_i \cdot P$
Log-Linear	$P = \alpha_0 + \sum \beta_i \ln z_i$	$\partial P / \partial z_i = \beta_i / z_i$
Double-Log	$\ln P = \alpha_0 + \sum \ln \beta_i z_i$	$\partial P / \partial z_i = \beta_i \cdot P / z_i$
Quadratic	$P = \alpha + \sum \beta_i z_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \delta_{ij} z_i z_j$	$\partial P / \partial z_i = \beta_i + 1/2 \sum_{j \neq i} \delta_{ij} z_j + \delta_{ii} z_i$
Linear Box-Cox	$P^{(\theta)} = \alpha + \sum_{i=1}^{N} \beta_i z_i^{(\lambda)}$	$\partial P / \partial z_i = \beta_i z_i^{\lambda-1} P^{1-\theta}$
Quadratic Box-Cox	$P^{(\theta)} = \alpha + \sum_{i=1}^{N} \beta_i z_i^{(\lambda)} + \frac{1}{2}\sum_{i,j=1}^{N} \delta_{ij} z_i^{(\lambda)} z_j^{(\lambda)}$	$\partial P / \partial z_i = (\beta_i z_i^{\lambda-1} + \sum_{j=1}^{N} \delta_{ij} z_i^{\lambda-1} z_j^{(\lambda)}) P^{1-\theta}$

Table 2.1 List of possible hedonic price functional forms

2.3 Spatial Hedonic Price Model

Whenever we handle observation which is spatially organized in cross-section, we should consider spatial autocorrelation since the existence of spatial autocorrelation implies a lack of independence across observations and the estimates with ordinary least squares could be biased and inconsistent. Anselin and Bera (1998) explain the spatial autocorrelation and its importance as follows.

> "Spatial autocorrelation can be loosely defined as the coincidence of values similarity with locational similarity. … The existence of positive spatial autocorrelation implies that a sample contains less information than an uncorrelated counterpart. In order to properly carry out statistical inference, this loss of information must be explicitly acknowledged in estimation and diagnostics tests."

In the following sections, we go over the statistical tests to determine the appropriate spatial model and specification of spatial lag and error models.

2.3.1 Diagnostic Tests for Spatial Dependence

Multiple diagnostic tests for spatial dependence have developed. Among them are Moran's I test, Rao score (RS) test, likelihood ratio (LR) test, Wald test and Lagrange multiplier (LM) tests (Anselin and Bera, 1998). In this section, we introduce the robust LM test reported in Anselin *et.al* (1996) since this test is computationally simple and most robust among others. Robust LM test is based on OLS residuals, and is for spatial error autocorrelation in the presence of a spatially lagged dependent variable and for spatial lag dependence in the presence of spatial error autocorrelation (Anselin *et.al.*, 1996).

We now consider a spatial autoregressive model with a spatial autoregressive disturbance expressed as follows

$$\begin{aligned} P &= \rho W_1 P + Z\beta + u \\ u &= \lambda W_2 u + \varepsilon \\ \varepsilon &\sim N(0, \sigma^2 I) \end{aligned} \qquad (2.8)$$

where P is (N x 1) vector of observations recorded in N locations (e.g. housing prices in our case) , W_1 and W_2 are (N x N) spatial weight matrices, ρ and λ are the spatial parameters, Z is (N x k) matrix of independent variables and β is (N x 1) coefficients to be estimated. Spatial weight matrices represent degree of potential interaction between neighboring locations. Various types of specifications for the weight matrices are discussed in detail in the following section.

We are going to consider two types of tests, testing $H_0 : \lambda = 0$ in the presence of ρ and testing $H_0 : \rho = 0$ in the presence of λ for two cases where $W_1 \neq W_2$ and $W_1 = W_2$. The first case, a robust LM test for $H_0 : \lambda = 0$ can be expressed as

$$LM_\lambda^* = \frac{[\tilde{u}'W_2\tilde{u}/\tilde{\sigma}^2 - T_{21}(N\tilde{J}_{\rho\cdot\beta})^{-1}\tilde{u}'W_1P/\tilde{\sigma}^2]^2}{T_{22} - (T_{21})^2(N\tilde{J}_{\rho\cdot\beta})^{-1}} \qquad (2.9)$$

where $\tilde{u} = P - Z\tilde{\beta}$ are the OLS residuals, $\tilde{\sigma}^2 = \tilde{u}'\tilde{u}/N$, $T_{ij} = tr[W_iW_j + W_i'W_j]$, $(N\tilde{J}_{\rho\cdot\beta})^{-1} = \tilde{\sigma}^2[(W_1Z\tilde{\beta})'M(W_1Z\tilde{\beta}) + T_{11}\tilde{\sigma}^2]^{-1}$ and $M = I - Z(Z'Z)^{-1}Z'$. For the case of W_1=W_2, (2.9) can be rewritten as

$$LM_\lambda^* = \frac{[\tilde{u}'W\tilde{u}/\tilde{\sigma}^2 - T(N\tilde{J}_{\rho\cdot\beta})^{-1}\tilde{u}'WP/\tilde{\sigma}^2]^2}{T[1 - T(N\tilde{J}_{\rho\cdot\beta})]^{-1}} \quad . \qquad (2.10)$$

As for the second type of the test, a robust LM test for $H_0 : \rho = 0$ is derived as

$$LM_{\rho}^{*} = \frac{[\tilde{u}'W_1P/\tilde{\sigma}^2 - T_{12}T_{22}^{-1}\tilde{u}'W_2\tilde{u}/\tilde{\sigma}^2]^2}{N\tilde{J}_{\rho\cdot\beta} - (T_{21})^2 T_{22}^{-1}} \tag{2.11}$$

for the case of $W_1 \neq W_2$, and for the case of $W_1=W_2$, it is

$$LM_{\rho}^{*} = \frac{[\tilde{u}'WP/\tilde{\sigma}^2 - \tilde{u}'W\tilde{u}/\tilde{\sigma}^2]^2}{N\tilde{J}_{\rho\cdot\beta} - T}. \tag{2.12}$$

Assuming $W_1=W_2$ is more realistic in practice since we can often expect the structure of spatial dependence to be the same for both the dependent autoregressive variable and the error term (Anselin et.al.,1996). These tests are tested against $\chi^2(1)$.

2.3.2 Spatial Weight Matrices

Tobler's (1979) "first law of geography" states that "everything is related to everything else, but close things more so". The question here is "how close is close?", or "how far is far enough to have no relation?" We have to determine a relevant "neighborhood set" indicating which locations have interaction and which are not. This is done by defining spatial weights matrix.

A spatial weights matrix is a N by N positive and symmetric matrix. By convention, the diagonal elements of the weights matrix are set to zero. There are mainly two major ways to define the weight matrix. One is based on the judgment of whether house i and house j are neighbors or not. In this case, elements of the matrix is shown as $w_{ij} = 1$ when i and j are neighbors and $w_{ij} = 0$ otherwise. The matrix is often row-standardized as

$w_{ij}^s = w_{ij} / \sum_j w_{ij}$. Note that row-standardized matrix may not be symmetric. In general, we choose how many nearest neighbors to be considered "neighbors" and see which weight matrices fit the model the most. The other way to define the weight matrix is based on the actual distance between the houses. We usually set up the cut-off distance. For the houses within the cut-off point, inverse of the distance between two houses are computed and entered as the element of the weight. It is zero if a house lies beyond the cut-off distance from the base house. It can be defined as

$w_{ij} = 1/d_{ij}$ for $d_{ij} \le \delta, w_{ij} = 0$ for $d_{ij} \ge \delta.$ where d_{ij} is the distance between house i and j, and δ is a cutoff distance value.

2.3.3 Spatial Models

When the spatial parameter on the autoregressive regressor, ρ is tested significant while the spatial parameter on error term, λ is not, we choose spatial lag model for estimation. Spatial lag model is expressed as follows:

$$P = \rho WP + Z\beta + \varepsilon \qquad (2.13)$$

where ε is assumed to be a vector of independent and identically distributed (i.i.d) error terms. When spatial lag model is selected, we know that a housing price is explained partially by the neighboring observations. In other words, this model is capturing spillover effects of neighborhood. The modeler is interested in measuring the strength of the relationship and the "true" effect of the explanatory variables after removing the spatial autocorrelation effects. The weight matrix is constructed to reflect the structure of

potential spatial interactions among observations (Kim *et. al*. (2003)). When the spatial autoregressive parameter, ρ is tested to be significant, ordinary least square (OLS) estimates are biased and inconsistent (Kelejian and Prucha (1998)). The spatial lag term $(WP)_i$ is always correlated with the error term since the term acts like endogenous variable. Furthermore, the spatial lag for the location i is correlated with the error term at i as well as the error terms at all other locations included.

Rewrite (2.13) as

$$P = (I - \rho W)^{-1} Z\beta + (I - \rho W)^{-1}\varepsilon . \qquad (2.14)$$

Since $(I - \rho W)^{-1}$ yields an infinite series $(I + \rho W + \rho^2 W^2 + \rho^3 W^3 + \cdots)$, orthogonality condition for OLS cannot be met as

$$E[(WP)_i \varepsilon_i] = E[\{W(I - \rho W)^{-1}\varepsilon\}_i \varepsilon_i] \neq 0 \qquad (2.15)$$

Estimating OLS by ignoring the spatial lag term when the term is relevant to the model causes the same problem as omitted variable. Suppose that a correct model is (2.13), but we estimate $P = Z\beta + \varepsilon$ by using OLS. We can see the estimated result is biased as follows.

$$\begin{aligned} b_1 &= (Z'Z)^{-1} Z'P \\ &= (Z'Z)^{-1} Z'(\rho WP + Z\beta + \varepsilon) \\ &= \beta + \rho(Z'Z)^{-1} Z'WP + (Z'Z)^{-1} Z'\varepsilon \\ E(b_1) &= \beta + \rho(Z'Z)^{-1} Z'WP \neq \beta. \end{aligned}$$

Given the bias and inconsistency, we have to use maximum likelihood estimation or instrumental variables estimation for this model (Anselin (1988), Kelijian and Prucha (1998), Kelijian and Prucha (1999)).

When ρ is tested insignificant and λ is significantly different from zero, we employ spatial error model which is expressed as follows:

$$\begin{aligned} P &= Z\beta + u \\ u &= \lambda Mu + \varepsilon \end{aligned} \qquad (2.16)$$

where ε is an N x 1 vector assumed to be distributed i.i.d. normal. The housing price is a function of the omitted variables at neighboring location as well as the independent variables. This model is appropriate when there is no theoretical or apparent spatial interaction between any house and its neighboring observations and the modeler is interested only in correcting the potentially biasing influence of spatial autocorrelation by using data with spatial features. OLS estimates are unbiased, but inefficient (Kim *et. al*.,2003). To see this, plug the second equation of (2.16) into the first and gain

$$P = Z\beta + (I - \lambda M)^{-1}\varepsilon \qquad (2.17)$$

The error covariance can be derived as

$$E(\varepsilon\varepsilon') = \sigma^2[(I - \lambda M)'(I - \lambda M)]^{-1} \quad . \qquad (2.18)$$

Therefore, it leads to nonzero error covariance between every pairs of observations (Anselin and Bera, 1996).

2.4 Second Stage Hedonic Method

The second stage of hedonic method is to determine demand function for each characteristic of the differentiated good. Deriving the demand function enables us to calculate welfare measures for the non-marginal change in characteristics. In order to estimate one demand function instead of one for each individual, we include individual socio-economic characteristics in this stage of estimation.

2.4.1 Theory

As we stated in section 2.2.1, demand function b_{ij} in Figure 2.2 cannot be derived from the first stage since all we can obtain is the information of the point where the implicit price function and the demand function intersects. In order to overcome this demand identification problem, estimating multiple hedonic price functions in the first stage by using data from separate markets has been employed the most in practice. By deriving multiple hedonic price function, we can identify two or more points on marginal bid function b_{ij}. See Figure 2.3 for the illustration. The basic idea of this multi-market approach is "finding cases where individuals with the same preferences, income, and other traits face different marginal implicit prices" (Freeman (1993)).

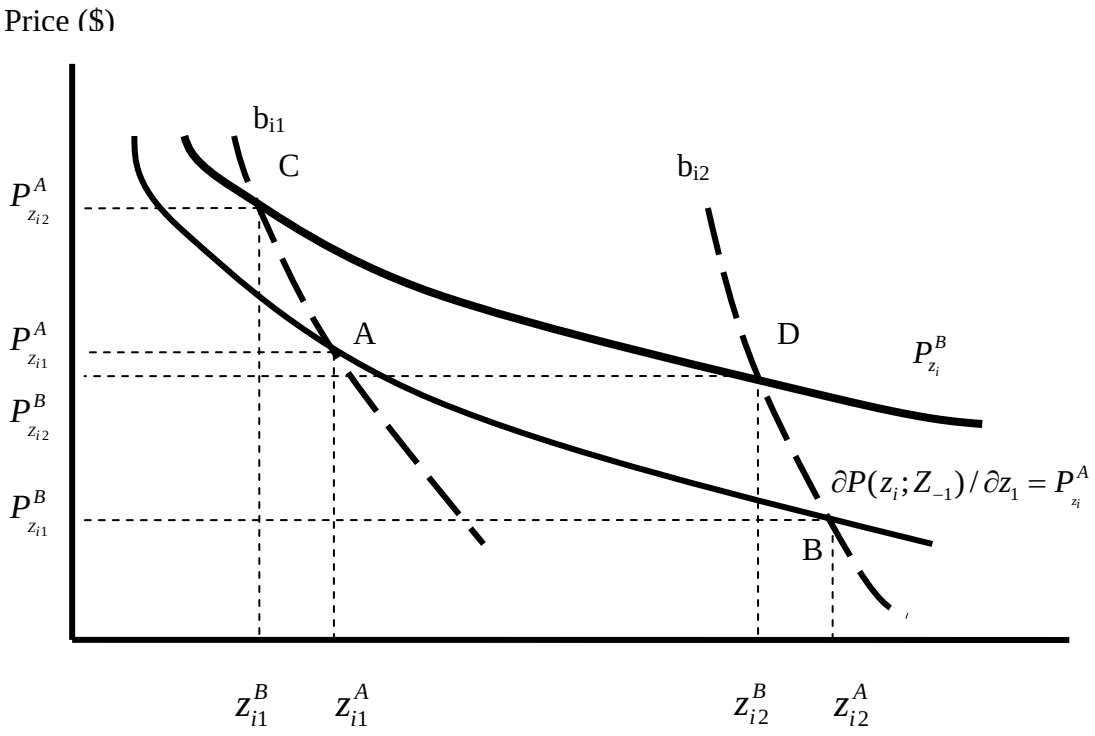


Figure 2.3. Identification of Marginal Bid Function

Figure 2.3 depicts the case of implicit price functions derived from hedonic price function derived from two separate markets. In addition to the point A and B determined for the first stage, now we can identify another points C and D on individual marginal bid functions. By controlling individual socio-economic characteristics in this stage, we can derive a single demand function in the end of the second stage estimation. Differences in hedonic price function from distinct markets arise from differences in the components of consumers and firms and their interactions. Varieties in supply could come from the differences in cost structures of firms while differences in demand could be from variations of the distribution of socio-economic characteristics among individuals within

a market. A critical assumption for connecting multiple points onto one marginal bid function is that "individuals with given vector of socio-economic characteristics have preferences over attributes that are identical across markets" (Taylor, 2003). Since supply structure is different in different markets, people with similar socio-economic profile are observed to make different house purchasing decision across different markets. When these assumptions are met and socio-economic characteristics are controlled properly, we can obtain demand function.

2.4.2 The Model

Prior to discuss the estimation procedure, we need to point out two sources of endogeneity problems. The first endogeneity arises from the fact that individuals choose the marginal price of the characteristics by choosing the quantity of the characteristics they demand for the case of non-linear marginal implicit prices. Choosing a point, say A in Figure 2.3, the consumer is choosing both marginal willingness to pay and the quantity of the attribute. This point is easily seen as the following log-linear hedonic price function and its marginal implicit price for characteristics i.

Hedonic Price Function: $P = \ln Z\beta + \varepsilon$
Marginal Implicit Price: $P_{Z_i} = \beta / z_i$

The second source of the endogeneity comes from the inclusion of an adjusted income since we need to linearize the budget constraint for the cases involving non-linear implicit prices. A budget constraint can be typically written as

$$Y = X + P(z_1, ..., z_n) \qquad (2.19)$$

where Y is the household income, X is expenditure for numeraire goods and $P(z_1,...,z_n)$ is the expenditure for the house with a set of attributes. When the budget constraint is non-linear, in order to derive the demand function analytically, it is necessary to linearize the budget constraint around the optimum point. The adjustment is implemented by first adding $\sum_{i=1}^{n} P_{Z_i} z_i$ to both sides of (2.19), then subtract $P(Z)$ from both sides, which is

$$Y^a \equiv Y + \sum_{i=1}^{n} P_{Z_i} z_i - P(Z) = X + \sum_{i=1}^{n} P_{Z_i} z_i \ . \qquad (2.20)$$

This adjusted income depends on non-constant marginal implicit price and again causes endogeneity.

Because of this endogeneity, two stage least squares estimation with instrumental variables is typically used in estimating the second-stage hedonic model. Estimates will be inconsistent if we ignore this endogeneity and estimate with OLS since "price depends on quantity and price is correlated with the error term in the equation explaining quantity demanded" (Palmquist (1991)). Here, the choice of proper instrument variables is very critical. These instruments should have the following properties, (1) correlated with the regressors, (2) uncorrelated with the error term, and (3) of full-rank (add new information) (Taylor (2003)).

Palmquist (1983) used age of the purchaser, dummy variable for the purchaser who is single, number of dependents in the family making the purchase and dummy variable for the purchaser who is black as instruments. Boyle *et.al* (1999) included purchaser's income, whether or not a property owner visited the lake before purchasing the property, whether or not the purchaser expected an improvement, decline, or no change in the

water clarity at the time the property was purchased, and whether or not friends or relatives of the purchaser also owned property on the lake at the time the property was purchased based on a survey. Bartik (1987) included a dummy variable for treatment groups since they used experimental data, as well as dummy variables indicating cities and time period. These studies utilized individual socio-economic data collected by individual surveys. Beron *et. al*. on the other hand used census tract level data. They included average household income net of housing expenditures and percentage of the population with a college degree.

In the remaining of this section, we illustrate the estimation procedure for this second stage. In order to estimate multiple hedonic price functions, we have to determine separate markets over geographic space and/or time. Suppose here that we identified two separate markets, A and B. We estimate two hedonic price functions for each market and obtain set of marginal implicit price functions for each market separately. For the case of log-linear hedonic price function, this process can be expressed as follows.

$$\begin{aligned} &\text{Hedonic Price Function for Market A: } P^A = \ln Z^A \beta^A + \varepsilon^A \\ &\text{Hedonic Price Function for Market B: } P^B = \ln Z^B \beta^B + \varepsilon^B \\ &\text{Marginal Implicit Price for Market A: } P_{Z_i^A} = \beta_i^A / z_i^A \\ &\text{Marginal Implicit Price for Market B: } P_{Z_i^B} = \beta_i^B / z_i^B \end{aligned} \tag{2.21}$$

We then pool data from both markets and include household socio-economic characteristics and market dummy variable into the data set. In the first step of the two stage least squares (2SLS) demand estimation, we estimate predicted value for the endogenous variable by using instrumental variables (IVs).

$$z_i = f(D^m, Y^a, \mathbf{C}) \tag{2.22}$$

where D^m is dummy variables for each market (one of them should be dropped), Y^a is the adjusted income (19), and **C** is a vector of socio-economic characteristics. Kahn and Lang (1988) and Beron *et.al*.(2003) included IVs as dummy variables for each market and its interaction terms with socio-economic variables while Palmquist (1984) "regressed the endogenous variables on all linear and quadratic terms in the exogenous socio-economic variables and a set of dummy variables for the urban areas".

In the second step of 2SLS, we actually estimate the demand function by using the predicted value estimated in the first step.

$$z_i = f(P_{Z_i}(\hat{z}_i), P_s(z_s), P_c(z_c), Y^a(P_{Z_i}(\hat{z}_i)), \mathbf{C}) \tag{2.23}$$

where $P_{Z_i}(\hat{z}_i)$ is the marginal implicit price evaluated at the predicted value $\hat{z}_i$, $P_s(z_s)$ and $P_c(z_c)$ are the marginal price for substitutes and complements of attribute z_i, respectively, $Y^a(P_{Z_i}(\hat{z}_i))$ is the adjusted income evaluated at the predicted value and **C** is the socio-economic characteristics. Once (2.23) is estimated, we evaluate all the variables except for z_i and P_{Z_i} at their mean values in order to derive the inverse demand function. Past studies listed in this section typically estimated (2.23) with linear, semi-log and Cobb-Douglas specifications.

The welfare change due to a change in the quantity of a variable can be measured by integrating under the estimated inverse demand function over the quantity change if the relocation is costly and the household decides to stay in the same house. It is given as

$$W_{z_i} = \int_{z_i^0}^{z_i^1} \frac{\partial P_{z_i}(z_i)}{\partial (z_i)} dz_i \quad . \qquad (2.24)$$

Since the estimated demand function is Marshallian demand, estimated welfare change is consumer surplus.

2.5 Conclusion

Hedonic method is composed of two parts. The first part is to estimate hedonic price function and compute implicit prices for variables determining housing prices. There have been many hedonic studies over past years. However, number of studies including water quality variables are much less comparing to the studies with air quality. Spatial lag or error model have been introduced in recent years and have been applied to hedonic price models. Together with the development of spatial econometrics model, the statistical tests such as robust LM test are developed as well to determine the appropriate spatial model.

The second part of the hedonic study is to estimate demand function and compute welfare change in non-marginal environmental quality change. This is typically done by using multiple hedonic price functions derived from separate markets and estimating by 2SLS with instrumental variables.

CHAPTER 3

DATA DESCRIPTION

3.1 General Data Description

In this chapter, we are going to discuss our data and variables included into hedonic price models. The preparation of general data other than water quality data are described in this section followed by the description of water quality data preparation in section 3.2.

3.1.1 Housing Data

We obtained Deed Transaction Data from 1985 to 1998 from Center for Urban and Regional Analysis (CURA). This data contains sales price of houses, sales date, address of the houses and other house characteristics, such as number of rooms, number of bedrooms, number of bath rooms, lot square footage and heat types as well as school district the houses belong. Single family occupations with less than twenty acres are included in our analysis. Deed transaction data is geocoded based on addresses by using roads centers file in order to determine the location of each house sold.

In our analysis, all prices (housing price and median household income) are discounted and expressed in 1996 dollar. Lot acreage, building square feet, number of bathrooms and garage square feet are included in the models as they are. Age of the house is derived by subtracting built year from year of the sale. Air-conditioning, deck and fireplace dummies are also included.

3.1.2 Neighborhood Data

School district ranking is obtained from Ohio Department of Education. Given the ranking within the state, we recalculated the ranking only for four counties (Erie, Lorain, Ottawa and Sandusky) included in the analysis. The map of school district and their ranking is shown in Figure 3.1. Median household income data is taken from census block group level data of year 2000.

3.1.3 Proximity Data

Proximities to the closest city and beach are calculated by using road-network, Arc Macro Language and ArcInfo. We compute two distances, from a house to the closet road network node and from the node to the destination, then add these two distances to gain total proximity to places. Distances to all the destinations are calculated and the smallest distance among them is adopted as the closest distance value.

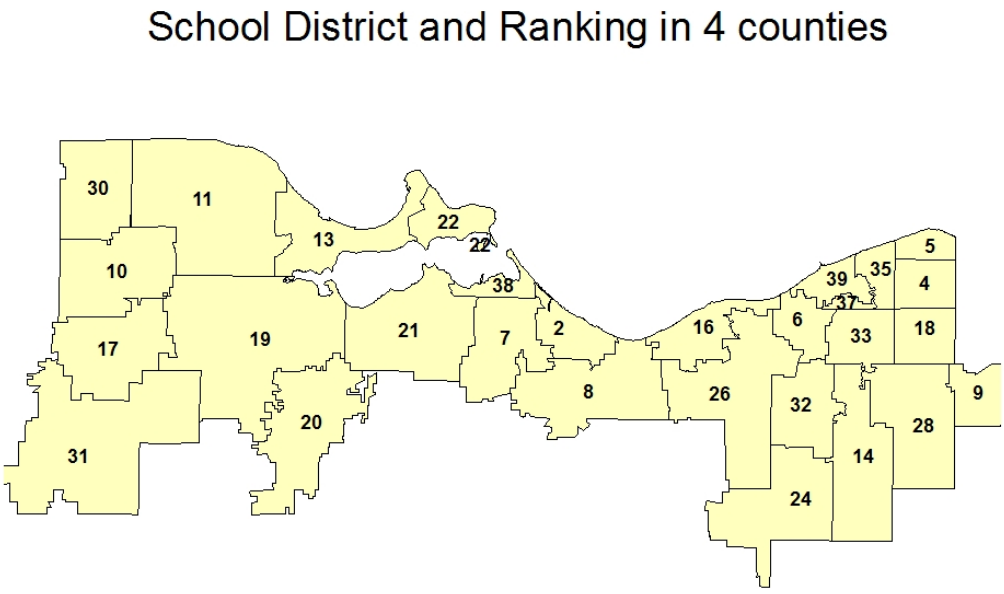


Figure 3.1 School District Boundaries and Ranking in Four Counties

3.2 Water Quality Data

Water quality measures in the hedonic study should be carefully chosen. The measures should reflect home purchasers water quality perceptions since individuals rely on what they can visually observe. In other words, measures typically used by scientists may not be a good representation of water quality to household purchasers. Fecal coliform count, E.coli bacteria content as well as secchi depth disk readings are used in our study as water quality measures.

3.2.1 Fecal Coliform Counts

Fecal coliform counts data has been obtained from Ohio Department of Health and Erie County Health Department. Fecal coliform and E.coli content are used as the indicator of beach closing. The associated standards can be found in Chapter 3745 of the Ohio Administrative Code (OAC) stating that the geometric mean E.coli content, based on not less than five samples within a thirty-day period shall not exceed 200 per 100 ml (for fecal coliform) or 126 per 100ml (for E.coli). Once the standard is exceeded, beach closing is posted. Measures of bacterial counts have been changed entirely from Fecal Coliform to E.coli in 1996 on all beaches since testing for E.coli bacteria may be the best method of analyzing such waters from organisms known to be harmful to humans.

Counties in Ohio switched measuring the bacterial counts from fecal coliform to E.coli in 1997. Since there is no direct conversion method available between those two measures, we have to handle data before and after year 1997 separately. The levels of these counts affect swimming and fishing activities directly and it is also hazardous to human health. Since we have housing data only up to 1998, we decide to include fecal coliform instead of E.coli and employ data from 1991 to 1996 in our analysis.

We first determine the closest beach to each house, and then assign Fecal coliform values of the beach which has the shortest distance from the house according to the year of the house purchase. Fecal coliform counts data have generally collected between May to September each year on the beaches along the Lake. After some trials with different ways of aggregation, annual average over one year before the purchase of each house has been adopted.

3.2.2 Secchi Disk Depth Readings

Secchi Depth Disk Readings Data is obtained from Stone Laboratory of Ohio State University and Sandusky Fisheries Research Station, Division of Wildlife, Ohio Department of Natural Resources. Secchi disk depth readings which indicates water clarity is also employed because it is a physical manifestation of the lake eutrophication and it is easily observed by individuals.

Secchi depth reading is an indicator of water clarity. The readings are taken between May and October in typical years. They are not taken from the same spots every month or year. Since the data varies over both space and time, using the raw data causes massive amount of missing observations. Therefore it is difficult to aggregate data over a year time period in order to assign data for each house depending on the sales date. In order to aggregate data meaningfully, we should take into account the structure of spatial autocorrelation of data. An example of secchi disk depth readings data are shown in Figure 3.2.

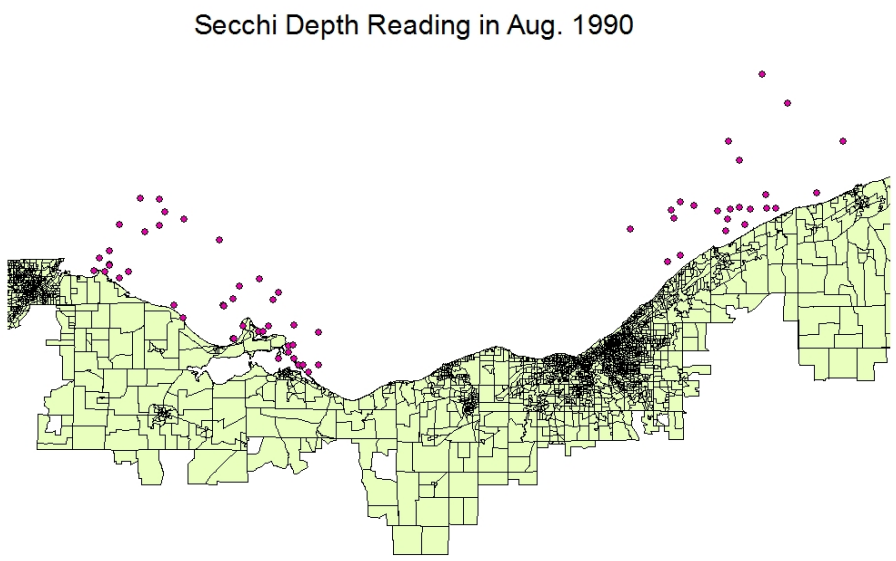


Figure 3.2 Locations of the Secchi Disk Depth Reading points in August 1990.

Empirical studies using water quality values as one of variables in the hedonic price function estimated did not either encounter or handle this problem. In general, multiple small-sized lakes are used to gain the variations in water quality. Therefore, researchers could use a single reading or index from a lake and assign the value to the surrounding houses of the lake. The examples of studies used multiple lakes are David (1968), Steinnes (1992), Michael, Boyle and Bouchard (1996), Feather (1992) and Boyle et.al. (1999).

Leggett and Bockstael (2001) used Inverse Distance Weight (IDW) interpolation method by using fecal coliform readings taken from monitoring stations along Chesapeake Bay. IDW is computed based on the assumption that things which are close to each other are more similar than those that are far away. The general formulation of IDW is

$$w(x,y) = \sum_{i=1}^{N} \lambda_i w_i, \quad \lambda_i = \frac{(\frac{1}{d_i})^p}{\sum_{k=1}^{N} (\frac{1}{d_k})^p}$$

where w(x,y) is the predicted value at location (x,y), N is the number of nearest negihbor points around (x,y), λ_i are the weights for each known point value w_i at location (x_i,y_i), d_i is Euclidean distances between (x_i,y_i) and (x,y), and p is the exponent which affects the weighting of w_i on w. We often see the case where p=2. Although they handle some spatial aspects of the environmental data, those data are taken from fixed monitoring station.

Some of recent hedonic studies with air quality as the environmental variables use kriging methods to handle the spatial variations of the variables. Anselin (2004) reported that estimated results differ depending on how researchers handle environmental quality data. He compared three different measures and interpolation methods (Thiessen Polygon which assign the value of closest monitoring station, Inverse Distance Weight and Kriging) and concluded that kriging is the best measure for the estimation.

Beron *et.al.* (2003) used kriging to deal with air quality data in four counties in Southern California. Yet, there is no hedonic study involving water quality used the kriging method as the measures for managing spatial environmental quality data.

Kriging is synonym to optimal prediction and is one of the important interpolation methods which provide a best linear unbiased predictor of any unobserved values. Kriging uses variograms as a weighting mechanism which assigns more influence to the nearer data points. Variogram is a measure of spatial variability and variogram distance measures the average degree of dissimilarity between a point to be estimated and a nearby known data value. Kriging makes inferences on unobserved values, takes into account the covariance structure as a function of distance and obtains best linear unbiased predictor.

We chose kriging to cope with spatio-temporal secchi disk depth readings data. All data collected for a year are plotted and then kriging was implemented over all of the data points. ArcMap was used for the implementation of kriging. An example of kriging is shown in Figure 3.3. Predicted values calculated with kriging are assigned to each beach location.

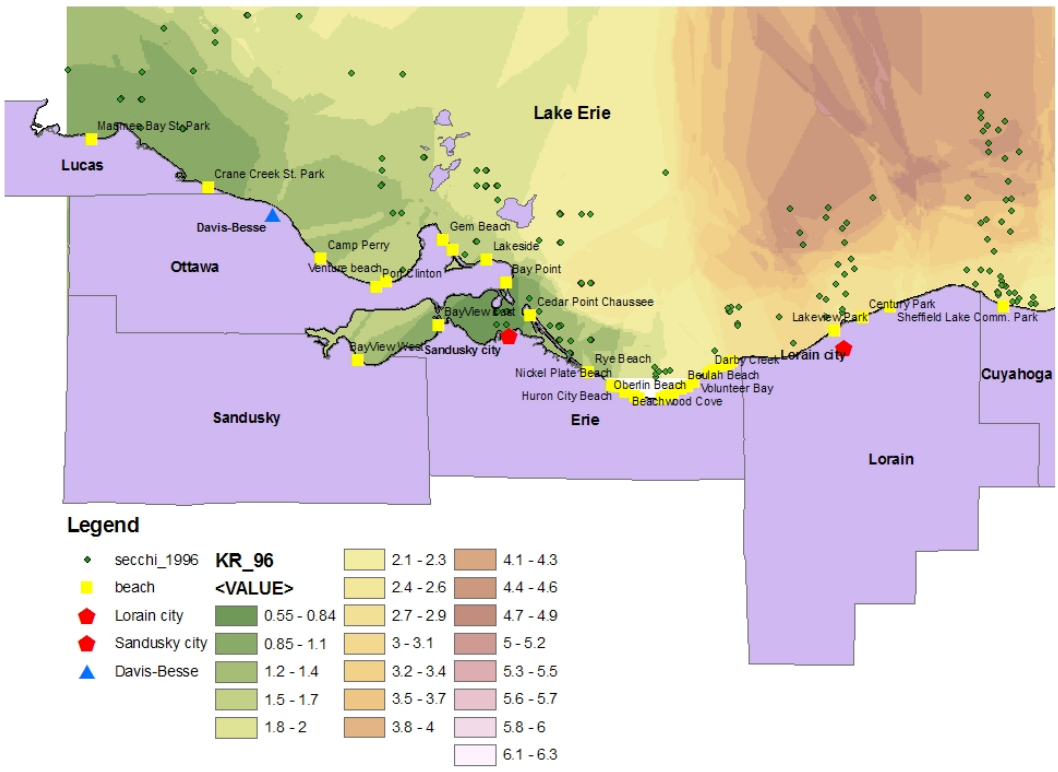


Figure 3.3 An Example of Kriging, 1996.

3.3 Descriptive Statistics of Data

Summary statistics of all the data used in clustering and first stage of estimation is listed in the table below. The first three variables are used in clustering and others are used in the first stage hedonic price model. Average housing price is 111,503 dollars in 1996 dollar. On average, the houses locate 5.8 km from the closest city, 9 km from the

Lake coast line and 12.6 km from the closest beach. Average age of houses is 30 years. The average fecal coliform counts house are facing is 255 counts per 100 ml while it is 2.2 meters for secchi disk depth readings.

Variable	Description	Unit	ALL Mean	St.Dev.	Min.	Max.
MedHHInc	Median Household Income, Census Block Group Data (2000)	1996$	36647	9480	4999	60499
CITY	Distance to the closest city	km	5.80	4.90	0.00	29.32
COAST	Distance to the closest coast line	km	9.05	7.27	0.00	40.75
DPRICE	Discounted housing price	1996$	111503	59186	50000	669292
LOTACR	Lot Acreage	acre	586.72	1806.78	10.00	78000.00
BLDGSF	Building Square Foot	sq.ft.	1649.75	607.49	196.00	5824.00
BATHN	Number of Bathrooms		1.42	0.56	1.00	5.00
GRGSQF	Garage Square Foot	sq.ft.	133.30	234.85	0.00	4040.00
AGE	Age of a House (Built year - year of purchase)	year	30.38	24.92	0.00	171.00
AIRCNDD	= 1 if there is an air-conditioning system		0.75	0.43	0.00	1.00
DECKD	= 1 if there is a deck		0.10	0.30	0.00	1.00
FIREPLD	= 1 if there is a fireplace		0.47	0.50	0.00	1.00
SDRANK	School district ranking within 4 counties		19.53	11.99	1.00	38.00
BEACH	Distance to the closest beach	km	12.56	8.68	0.01	48.13
FECAL	Fecal coliform counts	counts /100ml	255.99	281.44	12.00	2717.26
SECCHI	Secchi depth disk readings	meter	221.27	72.54	89.54	431.78
N			10655			

Table 3.1. Descriptive Statistics of Data

3.4 Conclusion

Many variables which are going to be used in our hedonic price estimation are prepared by using ArcGIS program based on raw data. The examples are the proximity variables derived by using Arc, Kriging used for secchi readings data, geocoding for housing addresses, and identification of census data and school district ranking for each data. Most of other variables are directly included into hedonic price estimation.

CHAPTER 4

CLUSTER ANALYSIS FOR SUBMARKET DETERMINATION

4.1 Submarket Definition

A typical definition of a submarket is “a set of dwellings that are reasonably close substitutes for one another, but relatively poor substitutes for dwellings in other submarkets (Grigsby et al., 1987).” More generally, “markets are truly separate if participants in one market do not consider houses in the other market when making purchase decisions (Taylor, 2003).” It is important to note that there is no contiguity requirement in the definition. As long as the houses in a submarket can be considered as close substitutes, it is possible to include non-contiguous houses or area into the same submarket.

Submarkets are typically defined in terms of geographical areas, physical characteristics of the dwellings, socio-economic characteristics of neighborhood, and in some cases defined by local real estate agents.

Examples of the geographical definition include the pre-existing geographical or political boundaries such as census block, postal code, school district or local political jurisdictions. Physical characteristics include housing structures and sizes (e.g., lot and floor area, number of rooms) and dwelling type (e.g., detached versus attached).

In more recent years, Cluster Analysis combined with Factor Analysis or Principle Component Analysis (PCA) has been used to determine submarkets. While the approach mentioned above rely on the existing "boundaries" that are defined by a researcher, the approach with PCA and/or Cluster Analysis does not depend on a priori definition of geographical boundaries, but relies on the underlying structure of raw data and their combination. Detailed description of Cluster Analysis and the related literature using these techniques are discussed in the following sections.

4.2 Overview of Cluster Analysis

The objective of cluster analysis is to uncover groups of homogenous observations. Clustering or classification originate largely in the natural sciences such as biology and zoology in the form of taxonomy, and numerous techniques have been developed in the discipline. Over time, the techniques have been adopted and used widely in areas such as engineering, marketing, archaeology, psychiatry, anthropology (Everitt *et.al.* (2001)).

When a researcher implements a cluster analysis technique, he/she has to make decisions over three major factors, namely, the clustering algorithm, the clustering criterion and the dissimilarity measure. We introduce two clustering algorithms, four clustering criteria and five dissimilarity measures including four measures that we

introduce as alternatives to the widely used measure, the Euclidean distance. Cluster analysis is used in our study in order to reveal the underlying housing submarkets inherent in our housing sales data covering four adjacent counties in coastal Ohio.

4.2.1 Clustering Algorithms

Large number of algorithms for grouping observations into clusters based on their similarities exist in the literature. In the area of market segmentation, mainly two clustering algorithms are used, namely hierarchical clustering and k-means clustering. In this section, we review these two techniques.

4.2.1.1 K-means Clustering

K-means clustering is implemented by assigning k cluster seeds and proceeding to group all observations with respect to their similarities to one of the cluster seeds. The number of clusters generated is the same as the number of cluster seeds initially given. Therefore, k-means clustering requires a priori knowledge of the number of clusters to be formed. For example, Bourassa *et.al.*(2003) used k-means clustering by presetting the number of clusters to the sales groups which are identified by the appraisers.

Once the cluster seeds are chosen, all objects are assigned to one of the clusters according to their distances to the cluster seeds. As clustering proceeds, the definition of each cluster is updated using the mean value of the observations assigned to that cluster, and the objects are regrouped according to their distances to the new cluster means.

These grouping and mean updating steps are iterated until the changes in cluster assignments are significantly small. The k-means algorithm minimizes the sum-of-squared-errors between the observations and the clusters (represented using their means).

4.2.1.2 Hierarchical Clustering

The k-means algorithm produces a flat data description where the clusters are disjoint and are at the same level. In some applications, groups of patterns share some characteristics when looked at a particular level. Hierarchical clustering tries to capture these multi-level groupings using hierarchical representations rather than flat partitions.

Hierarchical clustering does not require a priori knowledge of the number of clusters. If we start from individual observations (agglomerative method), at each successive iteration, two groups with the shortest distance are merged together. In the end, the algorithm produces a single group with all observations. Based on the similarity values obtained during merging, we can draw a hierarchical tree (called dendrogram) to observe which objects/clusters are grouped together at which iteration. It is called divisive hierarchical clustering instead of agglomerative clustering if one starts from one big group containing all observations, then in each following step one of the groups is divided into two according to a predetermined distance measure. In the end, the algorithm yields n groups each containing a single observation. Agglomerative procedures are probably the most widely used among the hierarchical methods (Everitt et.al. (2001)). Therefore, our discussion here on will focus on agglomerative hierarchical clustering.

Hierarchical clustering process is usually visualized by tree diagram, called a dendrogram. There are mainly two decisions each researcher has to make. The first is to decide what the "optimal" number of clusters is, in other words, in which stage in a dendrogram we can determine the most meaningful clusters. The second decision is about the "distance measure" that is used to find which pair of groups should be merged.

4.2.2 Clustering Criteria

In this section, we review four clustering criteria used with the agglomerative clustering algorithms. A clustering criterion determines how similar or dissimilar two groups of objects are. The methods we review in this section are the most widely used methods in the literature.

4.2.2.1 Single Linkage

Single linkage, also known as the nearest-neighbor criterion, chooses the pair of groups to be merged according to the distance between the closest pair of individual objects from each group. Here, closest distance means the most similar. This process can be illustrated using Figure 4.1 more clearly. First, the distance between all pairs of observations, initially belonging to individual clusters, is computed. Second, the pair with the smallest distance, which is 1 and 2 in this illustration, is found and these two observations are merged. At this point, we have four clusters, {1 2}, {3}, {4} and {5}. Third, the pair of clusters having the smallest distance, which corresponds to objects 4 and 5, is found again and these clusters are merged. Now we have three clusters, {1 2},

{3}, {4,5}. Next, cluster pairs are considered again. The clusters with the shortest distance are found as {3} and {1,2} because the distance between the objects 1 and 3 is the smallest possible distance according to the single linkage criterion. The object 3 is merged with 1 and 2 and two clusters are obtained, {1,2,3} and {4,5}. Finally, these two clusters are merged using the distance between the objects 3 and 4 as the smallest distance according to the single linkage algorithm.

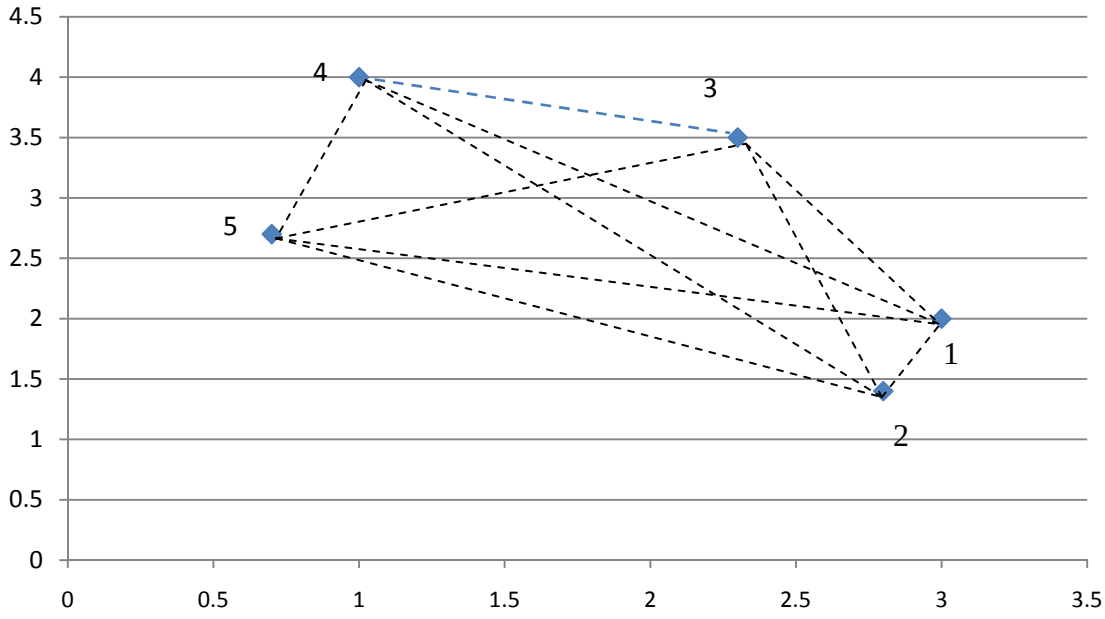


Figure 4.1 Example objects for illustration of Clustering Methods

4.2.2.2 Complete Linkage

Complete linkage, also known as the furthest neighbor criterion, is the opposite of single linkage in the sense that the distance between groups is defined by the most distant pair of objects. The most similar pair of clusters is merged as in the single linkage case, but the similarity is measured by minimizing the distance between the furthest objects in two clusters. We illustrate this process using Figure 4.1. First, 1 and 2 are merged, and second, 4 and 5 are merged as before. But now the distance between each pair of groups among {1 2}, {3}, {4 5} is measured using the furthest pair of objects. In other words, the distance between {1 2} and {3} is measured as the distance between 2 and 3, and the distance between {4 5} and {3} is measured as the distance between 5 and 3. Since the distance between 2 and 3 is shorter than the one for 5 and 3, {1 2} and 3 are merged together, forming the cluster {1 2 3}. Now the distance between {1 2 3} and {4 5} is measured as the distance between 2 and 4. Although the cluster formed in each step is the same in this setting both for single linkage and complete linkage, it is coincidental for our simple setting. Merging order can, and often do, differ between these two methods.

4.2.2.3 Group Average

Group average, also known as the unweighted pair-group method using arithmetic averages (UPGMA) treat the distance between two clusters as the average of the distance between each possible pairs of objects, one object from each cluster. For example, after grouping 1 and 2 in Figure 4.1, the distance between {1 2} and 3 is computed as

$$d_{\{1\ 2\}3} = \frac{1}{2}\ (d_{13} + d_{23})$$

where $d_{\{1\ 2\}3}$ is the distance between {1 2} and 3 and $d_{13}$and d_{23} are the distances between 1 – 3 and 2 – 3, respectively. Therefore, as a new object is merged into a cluster, the distances from that cluster to other clusters are updated.

4.2.2.4 Ward's Method

Ward's method, or Ward's minimum variance clustering method, is one of the most used clustering criteria after the group average criterion. Ward (1963) introduced a method, at each step of merging clusters, where two clusters whichever yield the smallest variance are merged together. In every step, variances for all possible combinations of clusters are computed and the combination corresponding to the smallest variance is chosen to be merged. In summary, the process follows three steps. First, the mean distance within each cluster is computed. Second, we compute the differences between each member in a cluster and its mean, square the difference and add them up within a cluster. Third, variances for all possible combinations of the tentative clusters are computed, and the ones returning the lowest variance are merged Romesburg (1984). The total within cluster sum of squares error, E is computed as

$$E = \sum_{g=1}^{G} E_g$$

$$E_g = \sum_{i=1}^{I} \sum_{k=1}^{K} (x_{gi,k} - \bar{x}_{g,k})^2$$

$$\bar{x}_{g,k} = (\frac{1}{n_g}) \sum_{i=1}^{I} x_{gi,k}$$

where g=1…G is the cluster index, i = 1, … I represents the object index within a cluster, k = 1…K is the k'th feature of an object, n_g is the number of objects within a cluster, $x_{gi,k}$ is the variable for the k'th feature of the i'th member within the g'th cluster.

Ward's method is commonly used in the existing studies which attempt to segment housing markets using cluster analysis (Bourassa et.al. (1999), Bates (2006)).

4.2.3 Distance Measures

How the similarity or dissimilarity between two objects with multiple attributes is defined is one of the most important issues to be considered in cluster analysis for grouping observations together. Euclidean distance is the most commonly used distance measure for continuous variables. For binary and categorical variables, there are other ways to define the distance since Euclidean distance does not represent the similarity well for these discrete variables in general. Furthermore, if the attributes of observations contain mixture of continuous and discrete variables, a hybrid distance measure for both types of variables has to be determined. In the following subsections, we review representative ways of defining distance for each type of variables.

4.2.3.1 Binary Variables

The distance for binary variables (or dummy variable as often used in econometrics models) in general is expressed as a "match" or a "mismatch". The general setting is shown in Table 4.1. In the table, **a**, **b**, **c** and **d** are considered as "matching scores" and are assigned according to the characteristics of the variable. For some cases, 1 – 1 match has the same meaning with 0 – 0 match, but for others may not be the same. An example of the former case could be gender in general sense. Both 1 – 1 and 0 – 0 match simply mean both entries have same gender. In this case, match score for **a** and **d** can be equal to one and zero for **b** and **c**. For the latter case, 0 – 0 match simply means the absence of certain characteristics and in some cases it does not contain useful information for determining the similarity between two objects. For example, the co-absence of wings in the context of taxonomy does not give enough information to define similarity between the two. In this case, score for **a** should differ from the one for **d**. Therefore, match scores have to be defined carefully depending on the type and the meaning of the binary variable.

		Individual i		
	Outcome	1	0	Total
Individual j	1	**a**	**b**	a + b
	0	**c**	**d**	c + d
	Total	a + c	b + d	p = a + b + c + d

Source: Everitt *et.al.* (2001) p.38

Table 4.1. Distance Definition for Binary Variables

4.2.3.2 Categorical Variables

Categorical variables contain more than two classes and the match score is typically assigned as one if a certain attribute fall into the same class for two objects and zero otherwise. Eye color can be an example of this case: match if two people have the same colored eyes, mismatch if they do not. If the classes for a categorical variable can be ordered numerically, it is possible to define different scores for mismatch cases by placing scores less than one and greater than zero for the cases which fall into adjacent classes. An example can be the income variable. We can assign one for the cases with same class of income, and a value between 0 and 1 to the cases which are not in the same class, but in adjacent classes. The magnitude of the "close match" cases has to be determined by reflecting the characteristics of the variable.

4.2.3.3 Continuous Variables

The most commonly used distance measure for continuous variables is the Euclidean distance which is defined as

$$d_{ij} = (\sum_{k=1}^{K} (x_{ik} - x_{jk})^2)^{1/2}$$

where $d_{ij}$is the distance between object i and object j, x_{ik} is the value of the *k*th attribute for object i and x_{jk} is the value of the *k*th attribute for object j. As Everitt et.al.(2001) reviews, there are other measures such as City block distance ($d_{ij} = \sum_{k=1}^{K} |x_{ik} - x_{jk}|$), more general Minkowski distance ($d_{ij} = (\sum_{k=1}^{K} |x_{ik} - x_{jk}|^{\frac{1}{r}} \quad (r \geq 1))$) and more.

4.2.3.4 Mixed Variables

In some cases, attributes which are used to determine the similarity contain both categorical and continuous variables. One possible way to handle mixed variables is to dichotomize all variables and use the similarity measure described for binary variables (Everitt *et.al.* (2001), Romesburg (1984)). The other way is to use Gower's general similarity measure although it is used very rarely in practice. This measure is given by

$$s_{ij} = \sum_{k=1}^{K} w_{ijk}\, s_{ijk} \Big/ \sum_{k=1}^{K} w_{ijk}$$

where s_{ijk} is the similarity between the *i*th and *j*th objects and w_{ijk} is set either as one or zero depending on whether the comparison is considered valid. For example, it is set to zero when the *k*th variable is binary and co-mismatch case can be excluded. For binary and categorical data, the similarity is set as one if two objects have the same value and zero otherwise. For continuous, Gower suggests to use the distance measure of

$$s_{ij} = 1 - \left|x_{ik} - x_{jk}\right| / R_k$$

where R_k is the range of the *k*th variable (Gower (1971), Everitt *et.al.* (2001)).

4.3 Literature Review on Cluster Analysis and Hedonic Price Models

Large amounts of research have been devoted to define meaningful submarkets over several decades. There are many different approaches introduced for the same purpose, from the approaches which use pre-determined boundaries such as political boundaries (census tract/block group), postal codes and school district to the use of Classification and Regression Trees (CART) (Clapp and Wang (2006)), hierarchical model (Goodman

& Thibodeau (1998)), latent variable analysis (Arguea and Hsiao (2000)), Principle Component Analysis (PCA) (Watkins (1999)), and Cluster Analysis (CA) (Goetzmann and Wachter (1995), Bourassa *et.al*. (1999), Bourassa *et.al.* (2003), Day (2003), Bates (2006)).

Before going into the reviews of individual studies, we would like to review Principle Component Analysis (PCA) briefly. PCA is used to reduce multidimensional data to a lower dimensional subspace and uncover the latent structure by constructing linear combination of variables from subset of original variables. It is often used to reduce the multicollinearity among multiple variables by creating factors by combining highly correlated variables.

In Bourassa *et. al.* (1999) study, both PCA and Cluster Analysis are used in order to determine submarkets in Sydney and Melbourne, Australia. They use two different data sets, one contains local government areas (LGA) and the other contains individual dwellings. Each data contain 43 LGAs in Sydney, 56 LGAs in Melbourne, 2307 individual dwellings for Sydney and 2354 houses for Melbourne. PCA is implemented by using twelve variables (distance to central business district (CBD), average number of bedrooms, percentage driving car to work, average number of cars, owner-occupation rate, distance to coast, population density, dwellings per km^2, median household income, percentage in public housing, percentage unemployed, distance to subcenter) for LGA data and eighteen variables are included for individual dwellings data (distance to CBD, percentage driving car to work, average number of bedrooms, percentage owner occupied, average number of cards, distance to coast, population density, dwellings per

km^2, median household income, percentage unemployed age of house, age of house squared, percentage in public housing, distance to subcenter, number of bedrooms, percentage detached, house value and number of problems). They identified three factors (linear combination of a subset of variables) for LGA data and six factors for individual dwellings data.

By including these factors, Cluster Analysis has been conducted. Both k-means and agglomerative hierarchical clustering are implemented. For hierarchical clustering, Ward's method is adopted as their clustering method. However, they do not mention the similarity measure used. As for k-means clustering, squared Euclidean distance is used. Number of clusters is set as five, same as a priori number of submarkets for both LGA and individual dwellings case. They also compared the output with the eight clusters case in order to observe which clusters are going to be divided into larger number of clusters.

In order to compare the outcome of clustering both from k-means and hierarchical clustering, they used weighted mean squared error (WMSE) (see section 4.6) computed from estimated hedonic equations for each cluster. They found that in most cases, k-means and hierarchical clustering have similar WMSE, but for one case (Melbourne individual dwelling data), k-means results are significantly better than the results from hierarchical clustering. In this study, they do not report the specification of the first stage hedonic equation or implement the second stage hedonic analysis.

Bates (2006) also uses PCA and Cluster Analysis to determine submarkets and compare clusters with existing planners' administrative boundaries in Philadelphia. Census block groups are used as the smallest data unit by commenting that "while it

would be preferable to have an even smaller unit of measurement, the block group is a relatively homogenous area that can capture the locational qualities of housing important to households." She uses thirty-one variables for PCA (% advanced math, % proficient math, % below-basic math, % advanced reading, % proficient reading, % below-basic reading, % detached houses, mortgage approval rates, medina income, % vacant land, % vacant residential, % to be demolished, % cleaned/sealed, % dangerous, % code violations, % LIHTC units, % public-housing units, % Section 8 units, % multifamily housing, % with bachelor's degree, % renters, % professional occupation, poverty rate, male unemployment rate, female unemployment rate, % female-headed households, % families on welfare, burglary rate, quality-of-life crime rate, drub-crime rate, car-theft rate). Seven factors are identified and used in clustering analysis. Hierarchical clustering with Ward's method is used as the clustering method. The similarity measure used is not mentioned in the paper. Although she identified six clusters, how the number of clusters have been determined is not clear.

With the identified clusters and 2000 house sales data, she estimated hedonic equation by regressing unit sale price on housing characteristics including building material, property type, size, area, stories, and garage, with and without submarket dummy variables. However, she does not proceed to second stage hedonic analysis given identified cluster information. She concludes that created submarkets better explain the variation in the housing market than preexisting planning-analysis sections (PAS).

The boundaries for identified submarkets and PAS differ greatly and each submarket includes area scattered in the whole study area. She concludes that "the PAS do not sufficiently define areas of housing that are relevant to household choices."

Day (2003) uses 3544 housing sales data of 1986 in Glasgow for his hedonic price analysis. The sales data come from 1027 different output areas (OAs) and further aggregated 38 postcode districts (PDs). He does factor analysis on a large number of neighborhood variables (25 attributes for OAs and 45 for PDs) and identifies six and four factors for OA and PD scale, respectively. He implements Cluster Analysis by including identified factors together with housing sales price, latitude and longitude of the house, proximity to the city center and structural characteristics such as dimensions, property type and property age. He uses the "hybrid" clustering method which combines partitioning and hierarchical clustering. First, partitioning method is used to generate 100 groups, and second, hierarchical method is used for further clustering. The reason for adopting this hybrid clustering is not described explicitly. However, the sentence "The drawback with these methods, however, is that they are computationally burdensome with large data sets" stated right above the introduction of hybrid method implies that the method is introduced in order to reduce the computational burden. Eight clusters have been identified by observing the shape of the dendrogram generated, and they are reduced to four clusters by merging two clusters which Chow test result indicates. He does not mention about either the similarity measure he used or the clustering linkage technique he adopted.

Spatial hedonic price function with spatial error specification has been estimated for each cluster by including 48 variables including 21 structural attributes, 10 factors generated for OA and PD, five accessibility attributes, and 12 environmental attributes including traffic noise and various views from the house (open land view, park view, industrial view, water view and so forth).

4.4 Discussion on Literature

In the studies reviewed in the previous section, in most of the cases the similarity measure used in the study is not even mentioned. Since it is not specifically discussed, we assume that they used Euclidean or standardized Euclidean distance which is often the default setting in clustering packages. Since Bourassa *et. al.* (1999) and Bates (2006) use factors generated from PCA directly into cluster analysis, their clustering variables are continuous. Day (2003) uses variables with different units for clustering including categorical variable for property type. Although he does not mention how he defines similarity measures, if he used Euclidean distance, he is adding up the differences of variables with different units. The use of standardized Euclidean distance solves this issue. However, since each variable is transformed into values with different ranges, the differences in value ranges can act as a weighting scheme for each variable without actually intending to do so when different variables are summed up to construct the similarity measure over multiple attributes. In addition, we are not sure how the categorical variable was treated in his study.

Similarity measure is an important “building block” for clustering. Clustering methods and techniques are all built on the similarity measure we choose. Therefore, depending on the types (continuous, binary, or categorical) of variables and characteristics (if categorical, ordered or not, if continuous, units) of variables, similarity measure should be chosen carefully.

4.5 Similarity Measures

In this section, we introduce similarity measures we suggest to use in our hedonic study. As we will discuss in the proceeding chapter, the variables we use for clustering include both continuous and categorical variables. Continuous variables are median household income, the proximity to the closest city and the proximity to coast line. Categorical variable is municipality variable including cities, villages and townships. In the effort of defining similarity between objects over multiple attributes with different units and variable types, we introduce four different measures we are going to use in our clustering analysis.

4.5.1 Euclidean Distance Revisited

Before we go into introducing our measures, we revisit Euclidean distance and show why this measure is not suitable for our case. If Euclidean distance is not standardized and the units of attributes are different, adding up the difference of one attribute to the other does not reflect the actual similarity between the object since the attributes with wider ranges will affect more to determine the similarity. In other words, the Euclidean

distance will implicitly assign more weighting to attributes with large ranges than those with small ranges. Moreover, in this way, we are adding up “apples” and “oranges” or physical distance (e.g. kilo meter) and income (e.g. dollar), for example.

Therefore, unitless standardized Euclidean distance is often used in order to avoid this problem. Standardization is done most commonly as follows.

$$Z_{ik} = \frac{X_{ik} - \bar{X}_k}{S_k}$$

where X_{ik} is the value of the *k*th attribute for the *i*th observation, $\bar{X}_k$ is the mean value of the *k*th attribute over *I* ($i = 0 \ldots I$) observations, S_k is the standard deviation of the values of the *k*th attribute. This standardization makes the variable’s distribution to have mean zero and variance of one. Note however that the range of the variable is not standardized into the [0,1] range.

It is not appropriate to use Euclidean distance in our case one is because of unit differences and the other is due to our mixed nature of the attributes (categorical and continuous). Moreover, even for the standardized Euclidean distance, the range is not [0,1] so that different variables can give different weights when we add up the similarity measure over different attributes. Therefore, it is ideal to have a continuous attribute falling into the [0,1] range especially when it has to be evaluated and added up together with binary or categorical variables.

Furthermore, standardization of Euclidean distance implicitly assumes normal distribution. However, the distributions of our clustering variables have log-normal like distribution, not normal. Therefore, the use of standardized Euclidean may not be appropriate. In the following subsections, we introduce the measure which could overcome these weaknesses.

4.5.2 CDF Transformation

We first introduce a transformation as an alternative to the standardization reviewed in the previous section for continuous variables. This transformation, that is based on the cumulative distribution function (CDF), has an advantage over the regular standardization because the transformed value ranges between zero and one.

Technically, it is described as follows. Given a random variable x with cumulative distribution function $F_x(x)$, the random variable $\tilde{x}$ resulting from the transformation $\tilde{x} = F_x(x)$ will be uniformly distributed in the [0,1] range (Papoulis (1991)). CDF transformation was used in the area of Pattern Recognition "to approximately equalize ranges of the attributes and make them have approximately the same effect in the computation of similarity between objects" (Aksoy (2001)). The concept of this transformation can be visualized as Figure 4.2. The motivation behind making the transformed variable have a uniform distribution in the [0,1] range is to make the values spread as much as possible in that range so that the discrimination ability of that attribute is increased.

The choice for the uniform distribution as a target for the transformed range comes from the fact that the uniform distribution on an interval is the maximum entropy distribution among all continuous distributions which are supported in that interval. Entropy is the amount of information contained in a random variable. An ideal attribute for identifying the similarity between objects is the one that has different values for different objects and similar values for similar objects. If there is no prior information about the distribution of the similarity, it is important to select attributes with lots of variation among items in order to distinguish different items better. For example, in order to define dissimilarity among multiple people, the attribute “the number of eyes” gives very little information about distinguishing one from the other. This kind of variable with very similar values for most of the items has very low entropy. On the other hand, the attributes such as height, weight, and age have higher entropy. Having maximum entropy is important because it ensures to describe the differences between objects as much as possible. If the range of the value [a b] is the only information given, the uniform distribution is the one that has the maximum entropy. Furthermore, this transformation does not assume any distributional forms. Therefore, we do not have any problem with using it on variables with non-normal distributions.

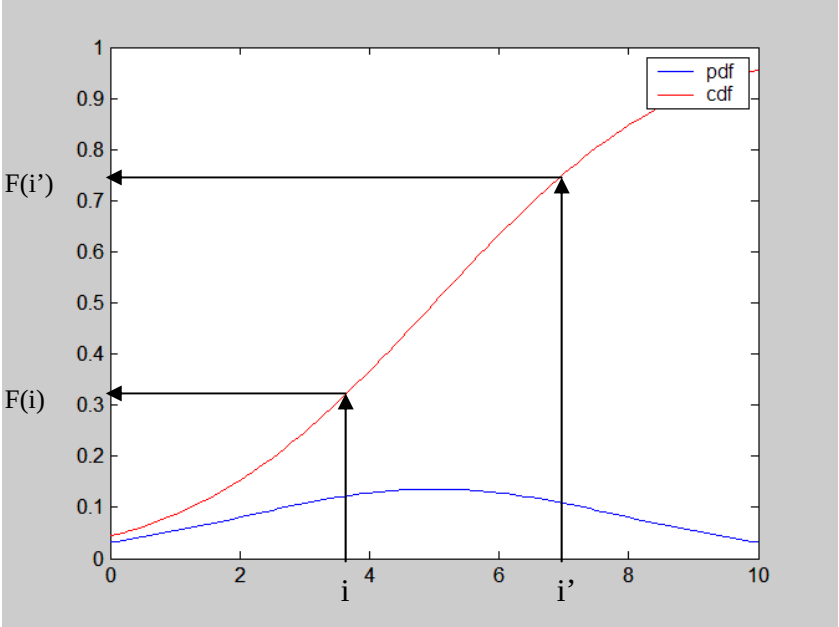


Figure 4.2 CDF Transformation

4.5.3 CDF + Hamming

Hamming Distance is typically used for binary data. Its basic logic is that if two observations have the same feature, a score of one is given to the pair, otherwise the score is zero as we described in Section 4.2.3.1. The overall distance between two objects is computed as the percentage of the matched counts.

In Section 4.2.3.4, we described the method to compute similarity measures by dichotomizing continuous variables and treating them as binary variables. Since dichotomizing may lose a lot of information about the attribute, we suggest to refine the measure, discretize instead of dichotomize continuous variables and treat them as

categorical variables. We first do the CDF transformation to make the variable fit in the range of [0 1] and then discretize continuous variables into ten bins with an increment of 0.1 (the increment is chosen empirically). Hamming distance is computed as giving 1 if the pair of the observations has the values in the same bin, otherwise 0, and taking the percentage of the counts over the whole occurrence. The score is determined as shown in Table 4.2.

The larger the numbers of bins are, the more the definition of similarity gets closer to the use of continuous variable. In the current study, we generated ten bins for all cases in order to compare the clustering results from four clustering practice with four different similarity measures with the same number of bins. We leave the determination of optimal number of bins for each variable as future work.

		Observation i									
	Bin	I	II	III	IV	V	VI	VII	VIII	VIIII	X
Observation j	I	1	0	0	0	0	0	0	0	0	0
	II	0	1	0	0	0	0	0	0	0	0
	III	0	0	1	0	0	0	0	0	0	0
	IV	0	0	0	1	0	0	0	0	0	0
	V	0	0	0	0	1	0	0	0	0	0
	VI	0	0	0	0	0	1	0	0	0	0
	VII	0	0	0	0	0	0	1	0	0	0
	VIII	0	0	0	0	0	0	0	1	0	0
	VIIII	0	0	0	0	0	0	0	0	1	0
	X	0	0	0	0	0	0	0	0	0	1

Table 4.2 Match Scores for Hamming Distance

4.5.4 CDF + Categorical 1

Categorical Method 1 is the applied measure of Hamming distance. Since we discretize continuous variables, generated categorical variables have a specific ordering. Therefore, those observations that have values in neighboring classes are more similar to each other than the ones with values in more distant classes. There is no rule of thumb for deciding how many neighbor classes should be included and how much score should be given. Therefore, we start out by giving half score for one-mismatch case (CDF + Categorical 1 method) and two thirds and one third to one-mismatch and two-mismatch cases, respectively, (CDF + Categorical 2 method) in order to see how much the change in definition matters.

After discretizing the variable into 10 classes, we give score 1 to the matching pairs and 0.5 to the pairs that are not matching exactly but the difference is just 1 neighboring class. By giving a partial score to the "not a match, but close" case, we attempt to include more information regarding the similarity between two observations from a continuous variable which otherwise could have been lost more in the dichotomization case. This idea is shown in Table 4.3. For example, if the *k*'th attribute for House *i* is in the class V and so as House *j*, the score will be 1. If the House *j*'s value is in the class VI while House *i* stays in the class V, then the score will be 0.5. Comparing to the CDF + Hamming case, we included more information about the continuous variable by including "not a match, but close" case in this way.

		Observation i									
	Bin	I	II	III	IV	V	VI	VII	VIII	VIIII	X
Observation j	I	1	0.5	0	0	0	0	0	0	0	0
	II	0.5	1	0.5	0	0	0	0	0	0	0
	III	0	0.5	1	0.5	0	0	0	0	0	0
	IV	0	0	0.5	1	0.5	0	0	0	0	0
	V	0	**0**	0	0.5	1	0.5	0	0	0	0
	VI	0	0	0	0	**0.5**	1	0.5	0	0	0
	VII	0	0	0	0	0	0.5	1	0.5	0	0
	VIII	0	0	0	0	0	0	0.5	1	0.5	0
	VIIII	0	0	0	0	0	0	0	0.5	1	0.5
	X	0	0	0	0	0	0	0	0	0.5	1

Table 4.3 Match Scores for Categorical Method 1

4.5.5 CDF + Categorical 2

Categorical 2 Method is the same as Categorical 1 Method except for the scoring scheme. Now we give 0.6 for the non-matching case with the distance of one bin and 0.3 for the non-matching case with the distance of two bins. This concept is shown in Table 4.4. These partial scores are given by attempting to reflect closeness of two observations with more information than simply giving the same score (zero) for mismatching cases.

		Observation i									
	Bin	I	II	III	IV	V	VI	VII	VIII	VIIII	X
Observation j	I	1	0.6	0.3	0	0	0	0	0	0	0
	II	0.6	1	0.6	0.3	0	0	0	0	0	0
	III	0.3	0.6	1	0.6	0.3	0	0	0	0	0
	IV	0	0.3	0.6	1	0.6	0.3	0	0	0	0
	V	0	**0**	0.3	0.6	1	0.6	0.3	0	0	0
	VI	0	0	0	0.3	**0.6**	1	0.6	0.3	0	0
	VII	0	0	0	0	0.3	0.6	1	0.6	0.3	0
	VIII	0	0	0	0	0	0.3	0.6	1	0.6	0.3
	VIIII	0	0	0	0	0	0	0.3	0.6	1	0.6
	X	0	0	0	0	0	0	0	0.3	0.6	1

Table 4.4 Match Scores for Categorical Method 2

4.6 Comparison of Clustering Methods

In order to compare the multiple clustering methods, we employ weighted mean squared errors (WMSE) from OLS conducted in each cluster as Bourassa et.al. (1999) compared k- means and Hierarchical clustering in their study. The smaller the WMSE is, the better the clustering method is for a given set of data. Weighted MSE is calculated as follows.

$$SE_U^2 = \frac{n_1 - k_1 - 1}{\sum (n_j - k_j - 1)} \cdot SE_1^2 + \frac{n_2 - k_2 - 1}{\sum (n_j - k_j - 1)} \cdot SE_2^2 + \cdots + \frac{n_s - k_s - 1}{\sum (n_j - k_j - 1)} \cdot SE_s^2$$

where SE_U^2 : variance of the unconstrained regression

$SE_1^2, ..., SE_s^2$: variances for the hedonic equations estimated for each of the subsamples

n_j : number of observations in the *j*th market

k_j : number of independent variables in the regression on the *j*th submarket

4.7 Determination of the Number of Clusters

Choosing the number of clusters in hierarchical clustering means choosing at which level the dendrogram should be cut. It is obvious that depending on the shape and the pattern of a dendrogram, the "best" number of clusters for given data differs. There are a few methods suggested to determine the number of clusters in different disciplines and this issue is still an undergoing research topic.

Since we use WMSE to choose the clustering method, by using the same criterion, we try to identify the "knee-point" by plotting WMSE. The knee-point is a point where the change in WMSE is small when the WMSE with a certain number of clusters is compared to the WMSE with one more cluster. When there is a tradeoff relationship between two variables (in our case, the number of clusters and WMSE), finding a knee point is a commonly used method (See for example Salvador and Chan (2004)).

4.8. Conclusion

In this chapter, components of Cluster Analysis have been discussed. Cluster Analysis consists of three major elements, clustering algorithms, clustering criteria and similarity measures. When we handle clustering variables with different units and types (continuous, binary or categorical), we have to pay attention to the choice of the similarity measures although existing studies seem to ignore the importance of this choice.

The use of Euclidean distance faces the problem of adding up variables with different units. In addition, Euclidean distance implicitly assigns more weighting to attributes with large ranges than those with small ranges. Standardized Euclidean distance solves this issue to some extent. However, it is not suitable for the case with mixed variables, and furthermore, it does not completely solve the issue of variables with different ranges since the standardization does not make variables to fall into equal ranges.

We introduced four different similarity measures in order to overcome these issues as much as possible. In the following chapter, we apply them to our hedonic study.

CHAPTER 5

APPLICATION OF CLUSTER ANALYSIS TO LAKE ERIE CASE

5.1 Data

We implemented two sets of Cluster Analysis, one is by using individual housing data directly and the other is by using census block group. In general, using the smallest unit possible (here, individual houses) as the minimum building block of the clustering is ideal since we do not have to assume anything about the underlying structure. Including census block group as the building block (smallest unit) in Cluster Analysis means that we assume the houses within the same census block group belong to the same housing submarket. The reason for the use of census block group is that it ensures the houses which share the same or very similar “neighborhood” are categorized into the same cluster. Separately, Bates (2006) used census block group as the smallest unit in her clustering since individual housing data was not available. Therefore, the use of census block group provides us with the comparison of the methods in terms of the assumption we have to make for the use of census block group and the effect of grouping geographical neighbors together.

The following variables are included as the "filter" for clustering in both cases.

- Median household income (census block group level)
- Distance to the closest city
- Distance to the Lake coast line
- X, Y coordinates
- Municipality (City, Village, Township)

The inclusion of the median household income is based on the household sorting theory. It is observed that households tend to sort themselves into neighborhoods with similar household income, education and race. Since it is often the case that income, education and race are highly correlated, we include household income in order to represent one of the main sorting factors. Distance to the closest city is included because monocentric model theorized that distance to city is important in terms of transportation costs to employment centers and shopping destinations. The theory assumes the tradeoff between the land price and transportation costs. Municipality is included based on the Tiebout model. Tiebout states that households sort themselves into neighborhoods according to a bundle of public goods and services provided by local municipalities. Distance to the Lake coast line is included based on the same theory since we consider that the amenities the Lake provide have gradient effects over the houses and differ for the houses close to the Lake and the ones away from the Lake. Especially because we are interested in the influence of water quality on housing prices, we included this variable in order to pick up the Lake's effect on the market segmentation. X, Y coordinates are included for two reasons. One is because in housing market, the physical location of a house is very

important, and the other is because these values become the sorting factor for collecting the neighborhood houses together in the same cluster. We are not going to include these variables into the first stage or second stage of hedonic analysis directly in order to avoid influencing the variability of each variable in the estimation stage.

Distance to the closest city is computed by using road network for 21 cities from individual houses or the centroids of the census block groups. Cities included for the distance computation are listed in Table 5.1 and city locations are shown in Figure 5.1.

Amherst	Clyde	North Olmsted	Port Clinton
Avon	Elyria	Northwood	Sandusky
Avon Lake	Fremont	Oberlin	Sheffield Lake
Bay Village	Huron	Olmsted Falls	Strongsville
Bellevue	Lorain	Oregon	Vermilion
			Westlake

Table 5.1. List of Cities Included for “Distance to the Closest City” Calculation

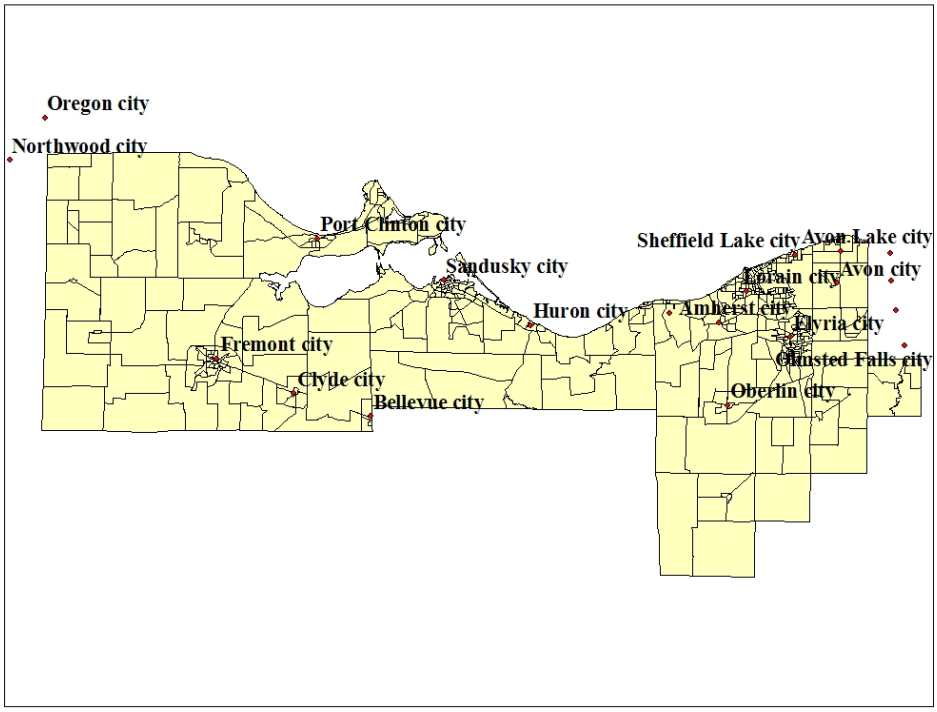


Figure 5.1 Location of Cities Included for "Distance to the Closest City" Calculation

Distance to the lake coast line is measured as the straight line distance between a house or a centroid point and the closest coast line. 211 distinguishable municipalities (townships, cities, villages) are included for the Municipality variable. Each observation is assigned to one municipality. The map identifying the boundaries of municipalities is found in Figure 5.2.

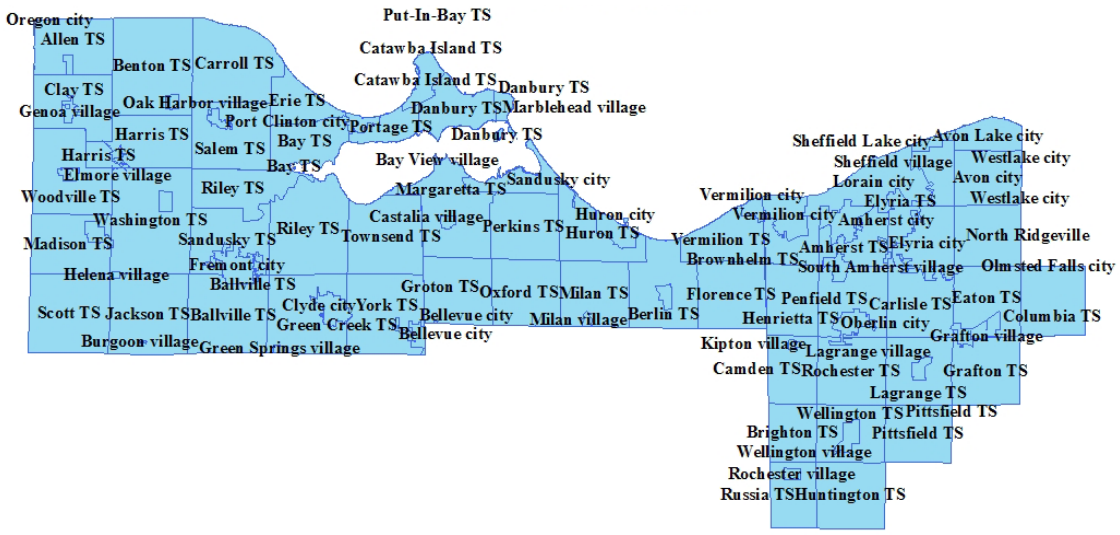


Figure 5.2 Cities, Villages and Townships Boundaries for Four Counties

Since we have both continuous (median household income, distance, x and y coordinates) and categorical (municipality) variables, it is important to pay attention to the choice of dissimilarity measures for the reason we discussed in Chapter 4. For both individual houses and the census block group case, we implemented four types of cluster analysis by using different dissimilarity measures explained in the previous chapter, namely, CDF transformation, CDF + Hamming, CDF + Categorical 1 and CDF + Categorical 2. Since median household income, distance to the closest city and distance to the coast line are continuous variables, we transform these as specified in each clustering method. X, Y coordinates are included by rescaling the variable into zero and one range without changing the distribution or relative magnitude. These coordinates are not quantized. Municipality variable is numbered from 1 through 211 and the similarity is coded in the Hamming way. In other words, it is set to one if the observations are in the same municipality and zero otherwise.

Variables used in clustering should reflect households' decision making process as well as the formation of submarkets. Although one may think that including as many variables as possible for clustering may help determine more realistic submarkets, it is not necessarily the case because of two reasons. One is because as the number of attributes increase in our clustering, the more "noise" we introduce in the clustering process. Furthermore, the distance computed between objects starts making less sense because many dissimilar objects could have very similar computed distances due to the cases such as "High value for A attribute + Low for B attribute" for object *i* and "Low for A, High for B" for object *j*. This is a commonly known problem, called "the curse of dimensionality" in the area of Pattern Recognition or Machine Learning in Engineering. The other reason for not including as many attributes as possible in clustering is related to the estimation of the hedonic price function. If we include the same variables used in the clustering into hedonic models, it is possible to cause the endogeneity problem. For example, if we use school district ranking as our clustering variable, the variations of this variable in each cluster are smaller than the case of not including it in the clustering. Therefore, if we include school district ranking in both clustering and the estimation of the hedonic price function, it will affect both the magnitude and the variance of the coefficient estimated. However, it is also true that the variables which determine the submarkets are the attributes considered in the housing purchases' decisions as well.

Therefore, it is important to choose the "right" amount of variables that are good representatives regarding the market segmentation for Cluster Analysis, and the variables used for clustering should be independent of the variables evaluated in the hedonic analysis, in our case, the water quality variables.

5.2. Clustering with Individual Houses

In this section, we report the clustering outcome and analysis for the case conducted by using individual houses data. Matlab is used for all clustering. Although there are built-in codes for Cluster Analysis, we added our own similarity measures into the codes. By implementing clustering with four different distance measures, we produced dendrograms as shown in Figure 5.3 – Figure 5.6. The dendrogram for CDF transformation looks different from the other three while the one for Categorical 1 and Categorical 2 appear quite similar to each other.

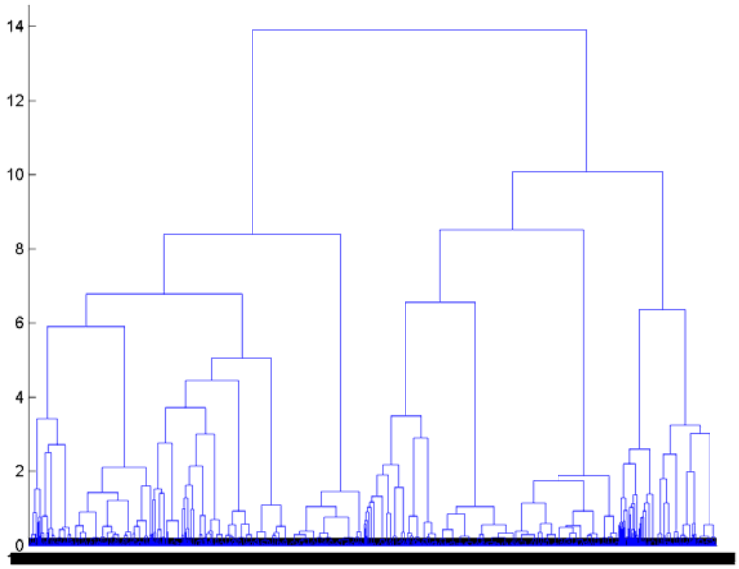


Figure 5.3. Dendrogram of Cluster Analysis with CDF Transformation: Individual Houses Case

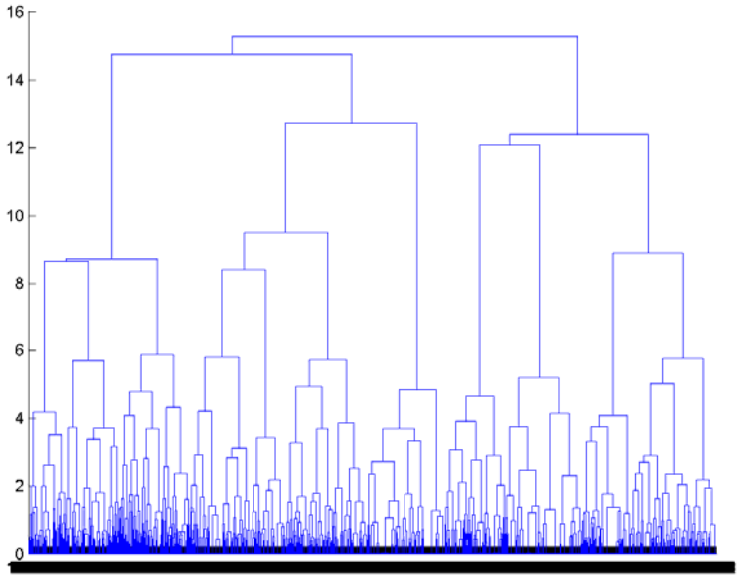


Figure 5.4 Dendrogram of Cluster Analysis with CDF + Hamming: Individual Houses Case

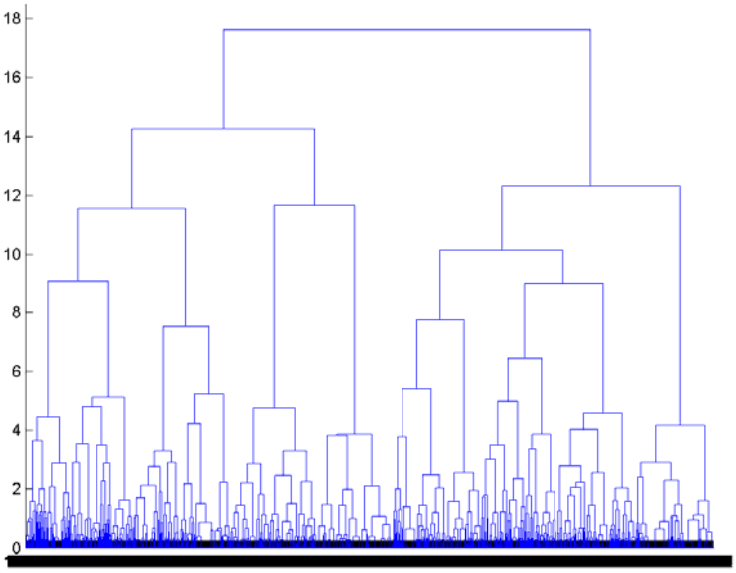


Figure 5.5 Dendrogram of Clustering Analysis with CDF + Categorical 1: Individual Houses Case

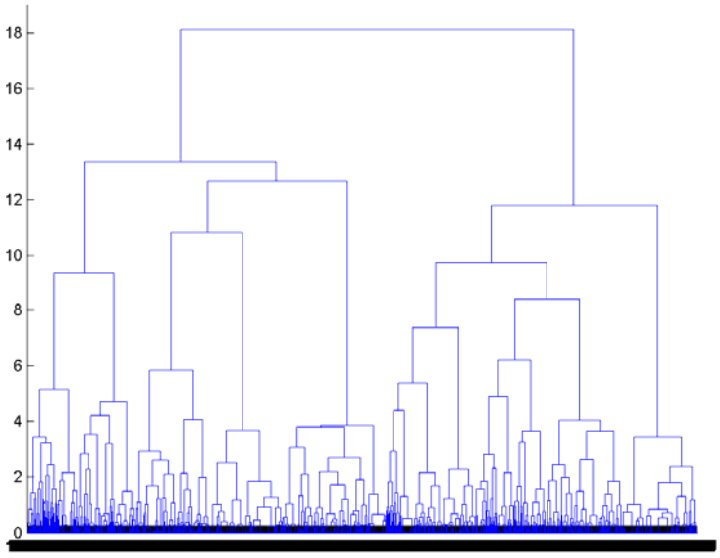


Figure 5.6 Dendrogram of Clustering Analysis with CDF + Categorical 2: Individual Houses Case

5.2.1 Comparison of Clustering Methods

In order to determine which clustering method works the best given the data, we compute weighted mean squared error (WMSE) for each clustering method. For each clustering method, we determine which observations are assigned to which cluster for the cases of number of clusters from one to twenty. OLS models which will be discussed in detail in the following chapter are estimated for each cluster, and WMSE is computed by using mean squared errors from OLS estimations for each cluster. Since number of clusters affects the magnitude of squared errors, we compare each method for the same number of clusters generated. WMSE for some number of clusters cannot be computed due to singularity of one or more variables. Calculated WMSE is listed in Table 5.2. Highlighted values are the minimum WMSE for a certain number of clusters created. By looking at the highlighted values, it is possible to say that Hamming method and Categorical 1 method have the smallest WMSE values for each cluster. By assuming that the number of clusters is greater than six, Categorical 1 method is a more likely candidate for the clustering given our housing data.

Once we choose the method of clustering, determining the optimal number of clusters is the next issue. As discussed in section 4.7 in chapter 4, we search the number of clusters by finding the "knee-point". Figure 5.7 plots WMSE values for the Categorical 1 method. In this figure, the knee-point can be identified around cluster numbers eleven and twelve. Weighted R-squared value indicates that eleven has better fit as shown in Table 5.3. Therefore, we choose to adopt Categorical 1 method with the number of clusters being equal to eleven for individual houses case.

N. Cluster	CDF	Hamming	Categorical 1	Categorical 2
1	600.22	600.22	600.22	600.22
2	285.45	313.13	283.03	357.34
3	232.82	183.23	211.32	338.31
4	206.47	147.24	156.09	196.39
5	145.04	113.37	146.55	192.56
6	93.12	104.26	119.08	115.50
7	82.73	87.60	81.95	112.17
8	77.31	80.09	73.65	102.20
9	64.87	57.63	56.52	96.06
10	54.65	51.88	52.12	93.58
11	48.09	47.19	46.67	93.19
12	44.08	42.22	41.64	70.08
13	40.92	40.32	39.91	59.73
14	38.62	37.91	38.74	49.88
15	36.94	34.96	n.a.	47.98
16	n.a.	33.82	n.a.	47.75
17	n.a.	31.81	n.a.	43.15
18	n.a.	30.26	n.a.	42.01
19	n.a.	28.90	n.a.	36.14
20	n.a.	26.67	n.a.	36.17

Table 5.2. WMSE Comparison: Individual Houses Case

	Categorical 1
1	0.697
2	0.679
3	0.691
4	0.673
5	0.666
6	0.659
7	0.646
8	0.651
9	0.652
10	0.653
11	0.654
12	0.652
13	0.649
14	0.650

Table 5.3. Calculated Weighted R-squares for Categorical 1 method

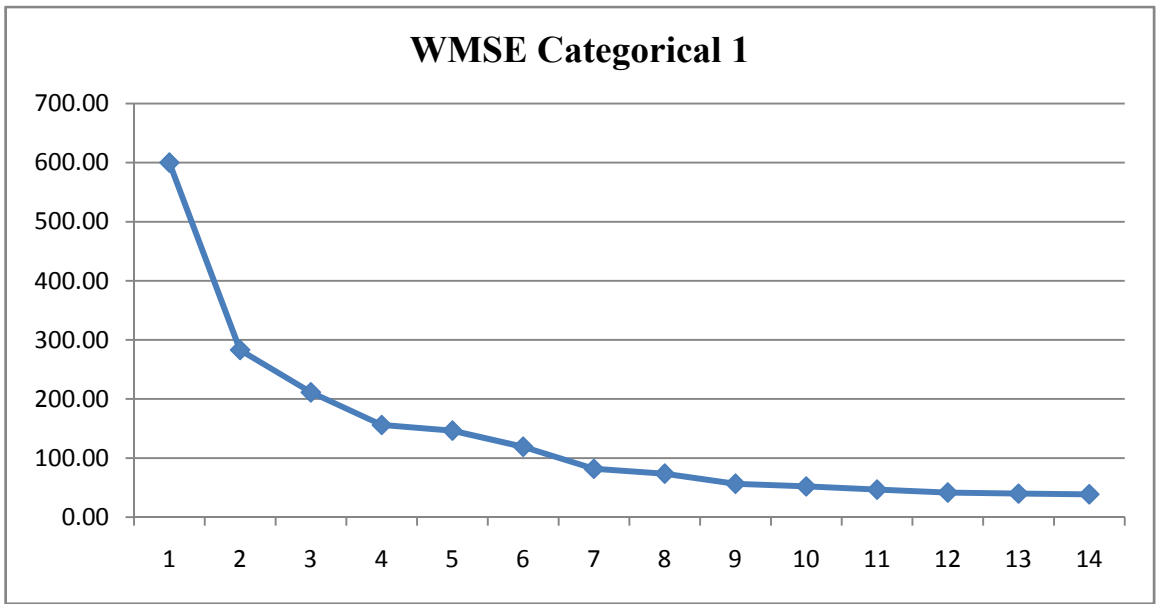


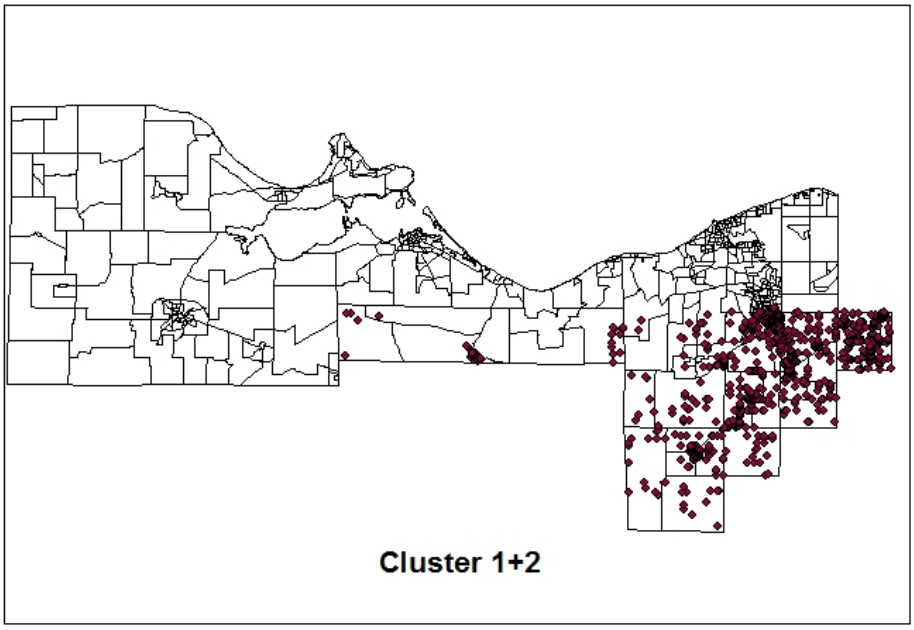
Figure 5.7 WMSE for CDF + Categorical 1 Clustering Method: Individual Houses Case

5.2.2 Analysis of Clustering Outcomes

By adopting Categorical 1 method with eleven clusters, the locations of the observations in each cluster are shown in Figure 5.8. Although there are some clusters with scattered observations located away from the main group(s), in general each cluster has one or more geographically distinguishable groups. Cluster 5, 6, 7 and 11 include lake shore observations although Cluster 7 contains a wide range of observations.

Table 5.4 lists the descriptive statistics for each cluster. The most distinguishable cluster is Cluster 11 which has the highest median household income, the highest discounted housing price, the biggest building square feet, the lowest age of houses, the highest average school district ranking and the shortest distance to the closest city. They

are facing the best water clarity, but the second worst fecal coliform level among other clusters. Cluster 1 has the largest lot acreage and houses in this cluster are located far from the coast line, and are also relatively far away from the closest cities. Houses in Cluster 3, 5, 8 and 11 are located within three kilo meter radius of the closest cities on average. Housing price is the highest for Cluster 11 and the lowest for Cluster 5 while lot acreage is also the lowest for Cluster 5. Houses facing the lowest fecal coliform counts and the lowest secchi depth readings are located in Cluster 8.



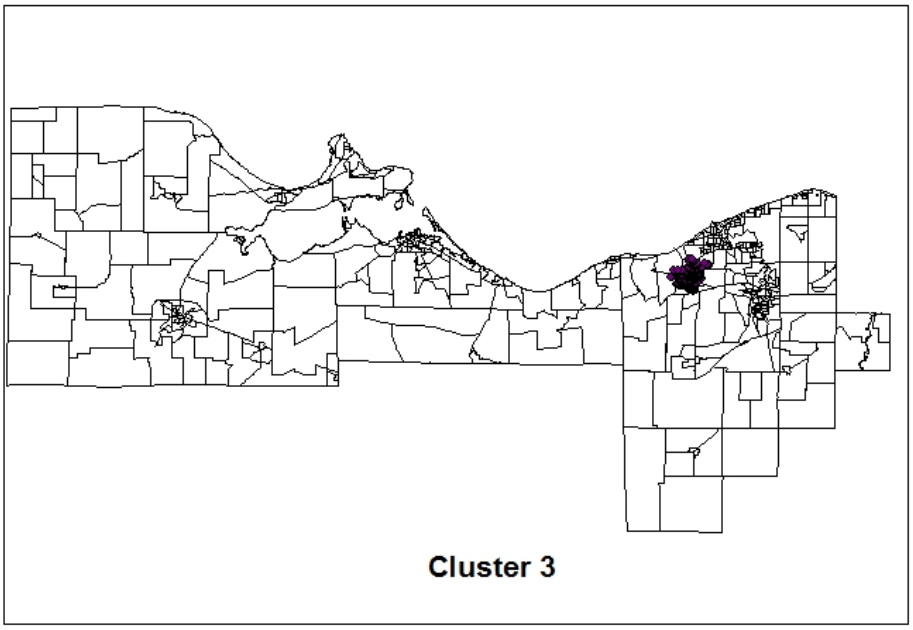
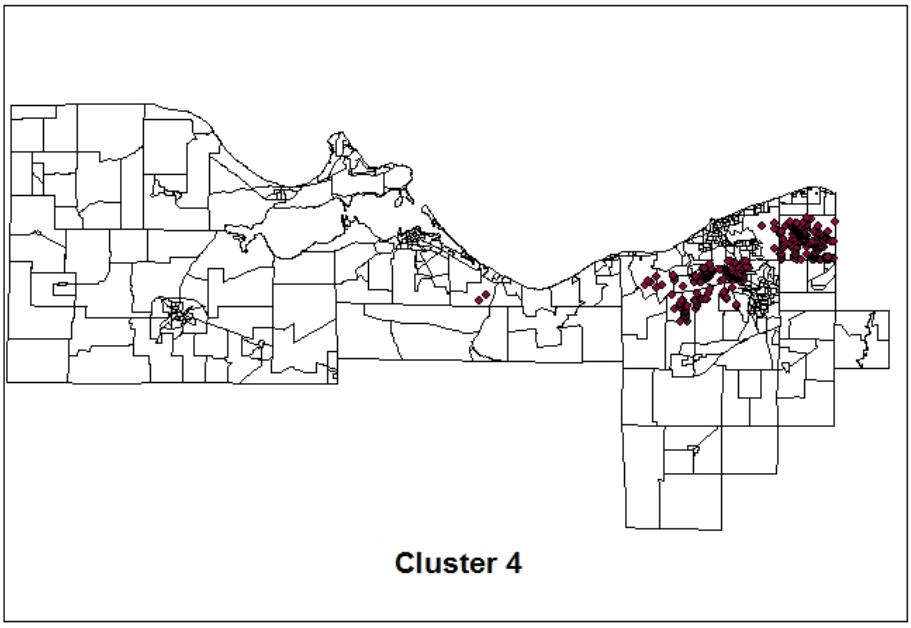


Figure 5.8 Observations in Each Cluster: Individual Houses Case

Figure 5.8 (continued)



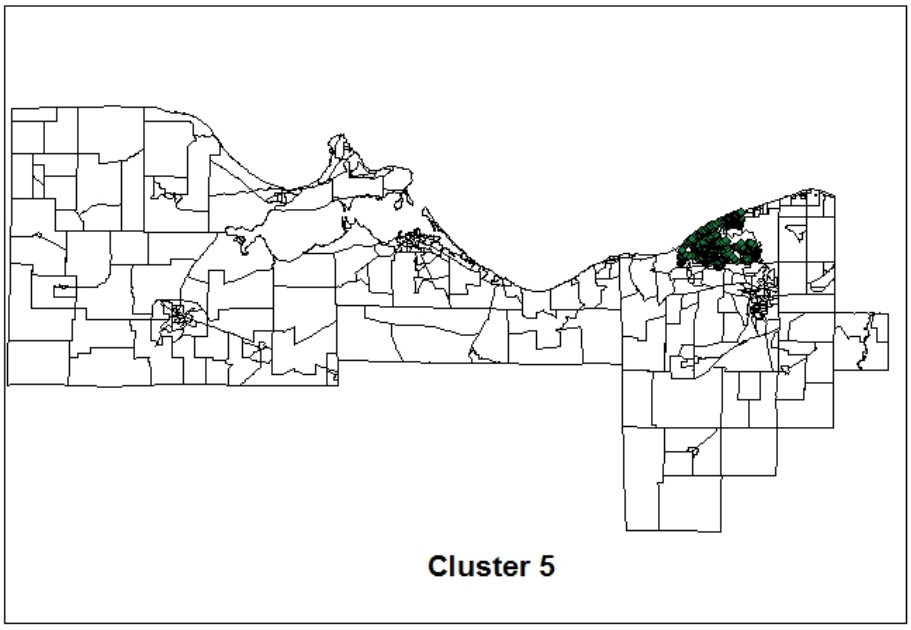
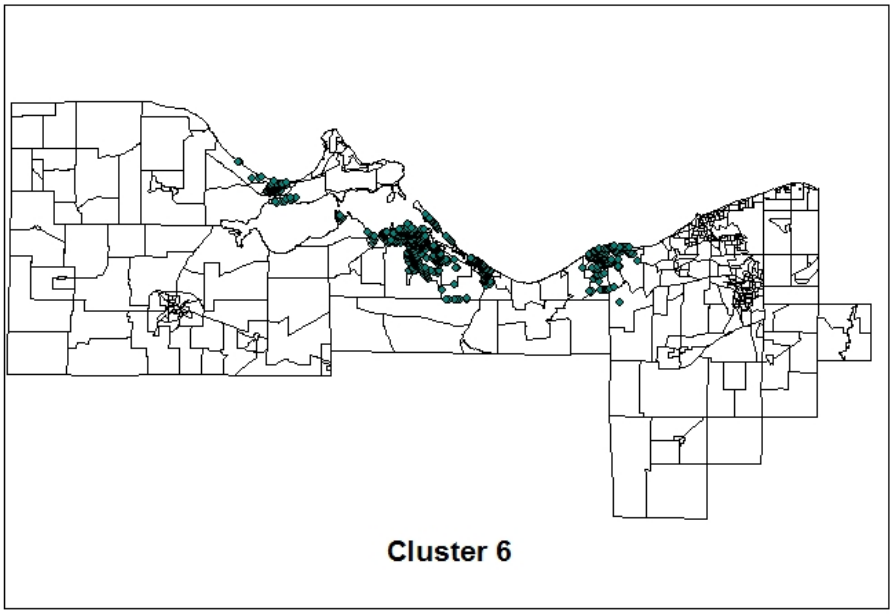


Figure 5.8 (continue)

Figure 5.8 (continued)



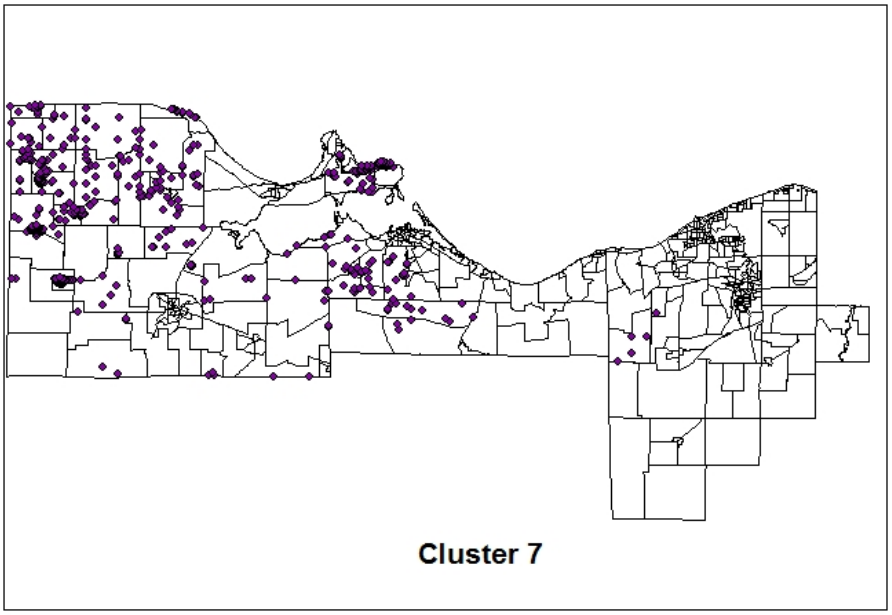
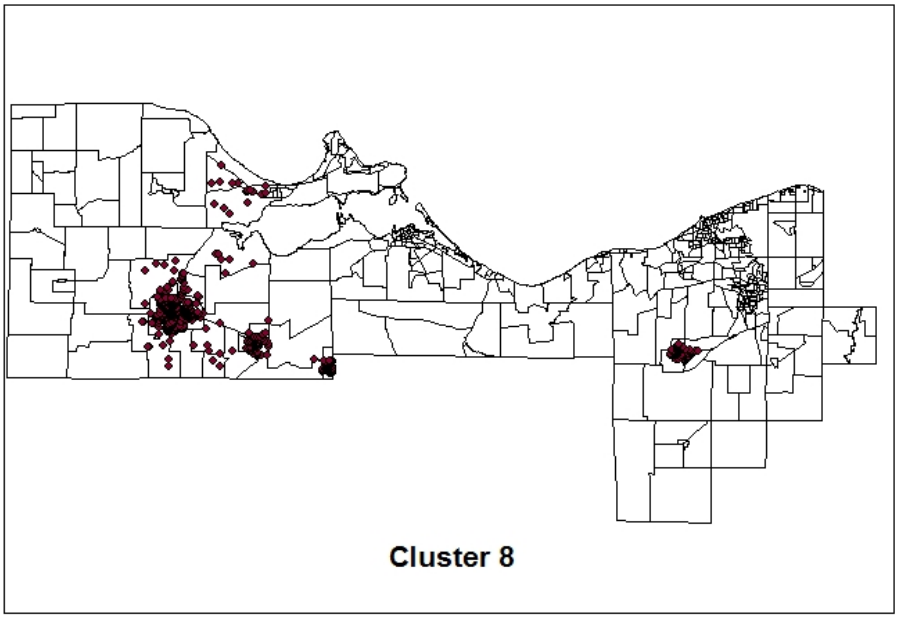


Figure 5.8 (continue)

Figure 5.8 (continued)



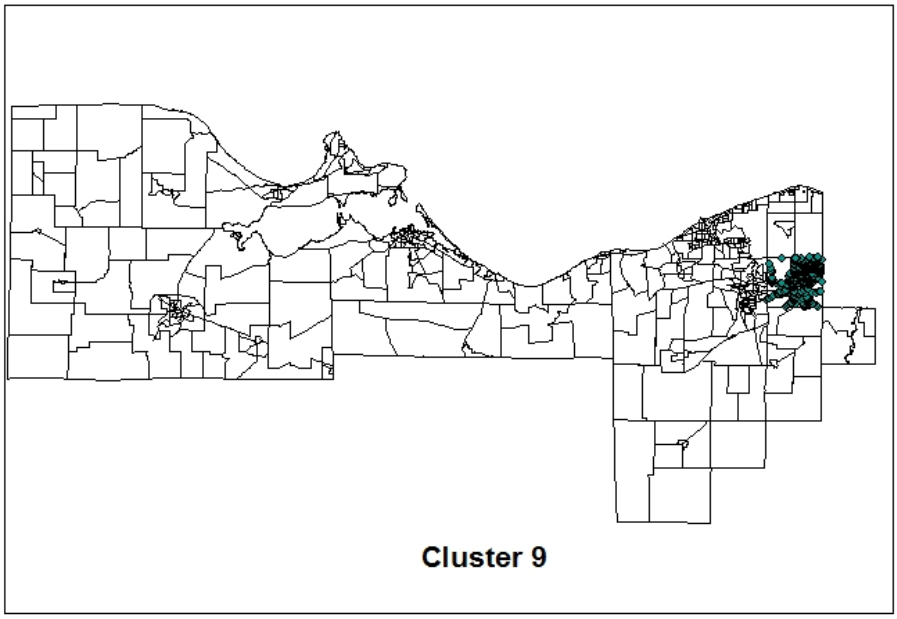
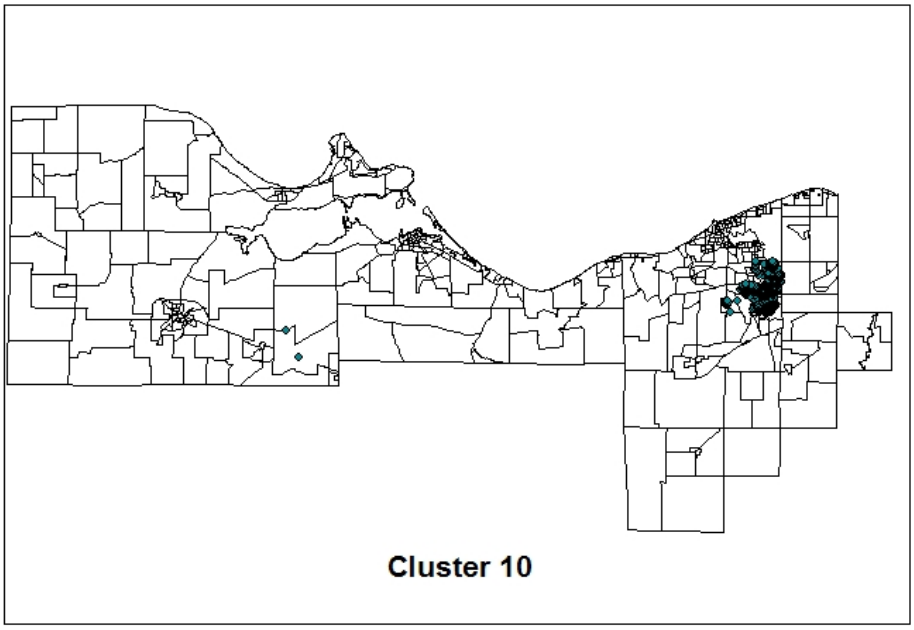
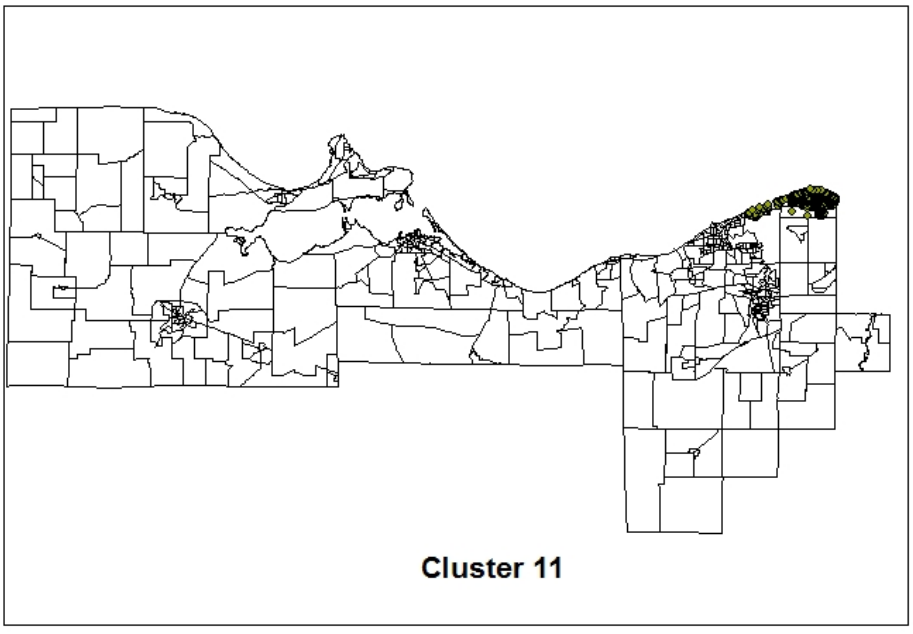


Figure 5.8 (continue)

Figure 5.8 (continued)





	Cluster 1				Cluster 2			
	Mean	Stand.Dev.	Min.	Max.	Mean	Stand.Dev.	Min.	Max.
MedHHInc	35607	5014	22857	43324	35116	6589	4999	43258
CITY	13.99	3.78	2.12	28.03	6.32	1.78	3.31	13.41
COAST	24.09	5.28	9.20	40.75	16.99	2.11	12.27	23.65
DPRICE	112300	49141	50195	465841	111357	43164	50195	406395
LOTACR	1752.61	2564.59	85.00	17800.00	1174.76	2290.50	122.00	18780.00
BLDGSF	1667.19	640.45	480.00	5506.00	1645.88	514.21	732.00	5190.00
BATHN	1.50	0.58	1.00	3.00	1.45	0.55	1.00	4.00
GRGSQF	83.34	246.91	0.00	4040.00	61.20	170.29	0.00	1032.00
AGE	22.60	23.77	0.00	171.00	20.71	18.36	0.00	95.00
AIRCNDD	0.88	0.33	0.00	1.00	0.89	0.32	0.00	1.00
DECKD	0.11	0.32	0.00	1.00	0.07	0.25	0.00	1.00
FIREPLD	0.48	0.50	0.00	1.00	0.53	0.50	0.00	1.00
SDRANK	21.88	7.16	5.00	36.00	21.49	10.73	1.00	33.00
BEACH	28.62	6.33	10.73	48.13	20.03	2.64	15.34	28.50
FECAL	252.60	239.50	29.28	2717.26	226.12	204.58	29.28	869.98
SECCHI	226.99	65.32	119.23	431.78	238.51	79.94	147.07	431.78
N		718				782		

	Cluster 3				Cluster 4			
	Mean	Stand.Dev.	Min.	Max.	Mean	Stand.Dev.	Min.	Max.
MedHHInc	36191	5993	30714	56859	40096	5800	14861	44706
CITY	1.94	0.78	0.02	4.10	4.32	1.66	0.00	7.48
COAST	4.54	0.93	2.31	6.69	6.82	1.54	3.30	12.14
DPRICE	132955	54728	50955	379900	129309	51635	50118	354922
LOTACR	419.93	651.84	40.00	13150.00	729.23	1549.29	117.00	19180.00
BLDGSF	1828.76	619.88	672.00	4288.00	1785.99	583.02	676.00	4796.00
BATHN	1.59	0.58	1.00	4.00	1.57	0.56	1.00	4.00
GRGSQF	9.52	67.79	0.00	576.00	33.82	129.29	0.00	1020.00
AGE	28.51	25.86	0.00	136.00	21.29	24.62	0.00	171.00
AIRCNDD	0.98	0.13	0.00	1.00	0.93	0.25	0.00	1.00
DECKD	0.13	0.34	0.00	1.00	0.07	0.26	0.00	1.00
FIREPLD	0.60	0.49	0.00	1.00	0.58	0.49	0.00	1.00
SDRANK	6.53	3.24	1.00	26.00	9.88	10.30	2.00	33.00
BEACH	9.43	1.36	6.12	12.19	9.49	1.95	3.82	16.48
FECAL	199.18	162.26	52.50	554.65	292.94	195.21	29.28	869.98
SECCHI	235.51	64.66	147.53	355.58	235.19	72.74	129.55	431.78
N		544				839		

Table 5.4. Descriptive Statistics for Each Cluster: Individual Houses Case

Table 5.4 (continued)

	Cluster 5				Cluster 6			
	Mean	Stand.Dev.	Min.	Max.	Mean	Stand.Dev.	Min.	Max.
MedHHInc	29383	8308	6419	56859	35584	9430	6461	60499
CITY	3.56	1.86	0.03	8.25	4.31	2.42	0.02	11.81
COAST	2.68	1.95	0.03	7.18	1.62	1.34	0.04	8.41
DPRICE	85840	35549	50000	311393	108182	62345	50000	582000
LOTACR	242.48	240.93	20.00	3920.00	374.51	1363.50	16.00	43000.00
BLDGSF	1458.15	483.72	640.00	4699.00	1563.22	598.03	196.00	4868.00
BATHN	1.27	0.49	1.00	4.00	1.31	0.51	1.00	4.00
GRGSQF	19.17	89.82	0.00	766.00	364.25	252.86	0.00	3651.00
AGE	33.00	17.16	0.00	94.00	36.90	24.99	0.00	164.00
AIRCNDD	0.96	0.21	0.00	1.00	0.37	0.48	0.00	1.00
DECKD	0.07	0.25	0.00	1.00	0.16	0.37	0.00	1.00
FIREPLD	0.35	0.48	0.00	1.00	0.37	0.48	0.00	1.00
SDRANK	33.63	10.63	6.00	38.00	16.97	12.38	2.00	37.00
BEACH	4.37	2.49	0.05	10.15	5.09	3.02	0.01	12.64
FECAL	216.99	176.17	29.28	869.98	203.53	462.58	18.24	2717.26
SECCHI	223.08	61.58	147.07	398.75	188.51	52.30	89.54	277.63
N		1229				1152		

	Cluster 7				Cluster 8			
	Mean	Stand.Dev.	Min.	Max.	Mean	Stand.Dev.	Min.	Max.
MedHHInc	33433	4179	16845	44034	29800	5445	13933	41528
CITY	18.16	5.39	3.94	29.32	2.46	1.85	0.06	13.04
COAST	12.16	8.90	0.04	31.37	13.72	2.95	0.97	20.31
DPRICE	95762	41709	50000	356236	86715	38941	50000	506000
LOTACR	1069.86	5037.60	21.00	78000.00	395.28	547.96	10.00	5030.00
BLDGSF	1494.98	481.06	204.00	3874.00	1532.40	524.57	202.00	4592.00
BATHN	1.26	0.46	1.00	4.00	1.21	0.43	1.00	5.00
GRGSQF	382.96	274.96	0.00	1730.00	349.62	253.87	0.00	1200.00
AGE	46.54	31.67	1.00	162.00	52.82	30.71	0.00	159.00
AIRCNDD	0.17	0.38	0.00	1.00	0.31	0.46	0.00	1.00
DECKD	0.14	0.35	0.00	1.00	0.13	0.33	0.00	1.00
FIREPLD	0.33	0.47	0.00	1.00	0.26	0.44	0.00	1.00
SDRANK	18.01	7.57	7.00	31.00	21.50	5.01	12.00	32.00
BEACH	22.00	13.15	0.06	43.79	19.34	3.59	2.69	27.67
FECAL	377.05	539.30	12.00	2717.26	87.26	99.82	14.15	867.92
SECCHI	182.59	44.88	89.54	258.80	178.50	53.25	118.96	355.58
N		693				971		

Table 5.4 (continue)

Table 5.4 (continued)

	Cluster 9				Cluster 10			
	Mean	Stand.Dev.	Min.	Max.	Mean	Stand.Dev.	Min.	Max.
MedHHInc	42040	3078	34444	46509	32516	6994	8347	48854
CITY	7.35	1.22	4.26	11.18	4.05	1.59	0.65	7.42
COAST	11.60	1.75	8.76	16.55	11.69	1.74	7.44	15.05
DPRICE	108561	30795	50000	332744	86877	31401	50000	334466
LOTACR	513.56	1109.75	53.00	15060.00	228.04	191.85	10.00	2769.00
BLDGSF	1682.51	422.32	538.00	4425.00	1457.09	516.98	576.00	5074.00
BATHN	1.49	0.54	1.00	4.00	1.29	0.52	1.00	5.00
GRGSQF	63.39	170.69	0.00	1200.00	60.35	154.41	0.00	835.00
AGE	19.63	15.72	0.00	170.00	33.89	17.98	0.00	128.00
AIRCNDD	0.92	0.28	0.00	1.00	0.89	0.32	0.00	1.00
DECKD	0.07	0.25	0.00	1.00	0.07	0.25	0.00	1.00
FIREPLD	0.55	0.50	0.00	1.00	0.37	0.48	0.00	1.00
SDRANK	16.58	3.92	6.00	26.00	32.02	2.58	14.00	33.00
BEACH	14.13	1.77	10.72	19.74	13.85	1.99	9.25	20.53
FECAL	258.54	188.61	67.25	869.98	310.07	249.32	14.15	869.98
SECCHI	243.13	83.72	147.89	431.78	229.11	70.93	142.49	398.75
N	1185				1334			

	Cluster 11			
	Mean	Stand.Dev.	Min.	Max.
MedHHInc	50968	9758	25720	60190
CITY	3.10	1.05	0.03	5.42
COAST	1.25	0.90	0.00	3.57
DPRICE	176869	95248	50609	669292
LOTACR	388.03	485.28	20.00	10000.00
BLDGSF	2104.02	825.27	646.00	5824.00
BATHN	1.74	0.61	1.00	5.00
GRGSQF	62.97	165.16	0.00	825.00
AGE	19.00	19.11	0.00	127.00
AIRCNDD	0.89	0.31	0.00	1.00
DECKD	0.12	0.33	0.00	1.00
FIREPLD	0.74	0.44	0.00	1.00
SDRANK	5.99	4.88	4.00	38.00
BEACH	3.41	1.58	0.04	6.00
FECAL	370.01	177.00	67.25	869.98
SECCHI	246.24	83.16	147.89	431.78
N	1218			

5.3 Clustering with Census Block Group

In this section, we report our clustering outcomes by using 417 centroids of census block groups. As we discussed in the beginning of this chapter, all the clustering variables are defined at or from the centroid of each census block group.

5.3.1 Comparison of Clustering Methods

As in the case with individual houses, we conducted four types of clustering by using different similarity measures. After clustering each case by using centroid data, individual houses data are assigned to the corresponding clusters, and estimated OLS results are used to compute the WMSE by using the same set of independent variables in the hedonic model. Generated dendrograms are shown in Figure 5.9 through 5.13.

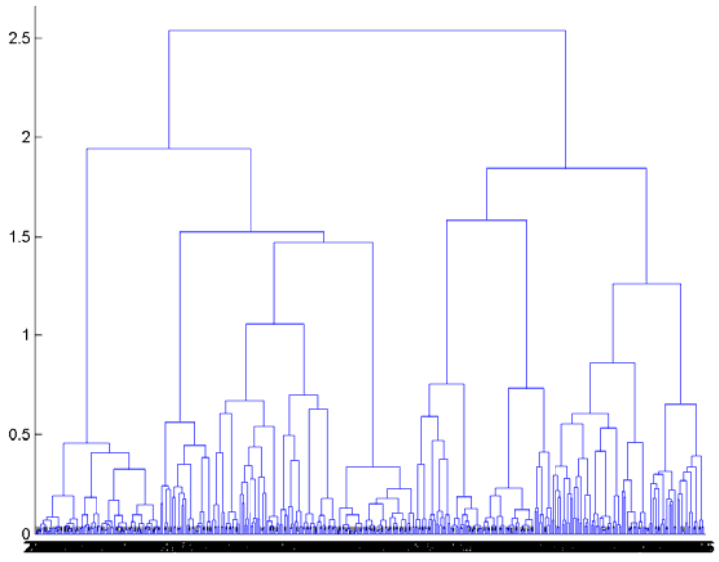


Figure 5.9. Dendrogram of Clustering Analysis with CDF Transformation: CBG Case

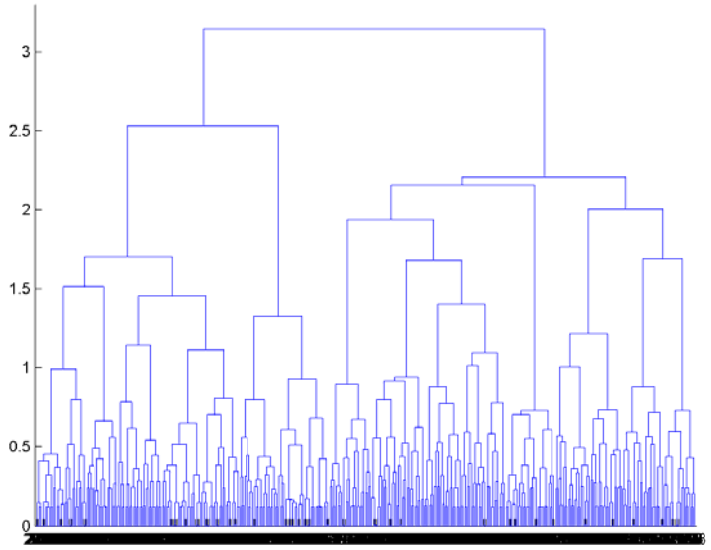


Figure 5.10 Dendrogram of Clustering Analysis with CDF + Hamming: CBG Case

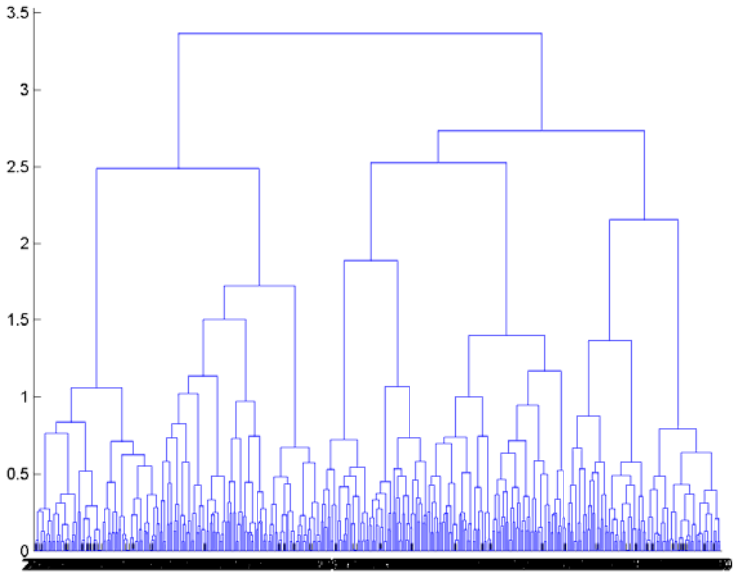


Figure 5.11 Dendrogram of Clustering Analysis with CDF + Categorical 1: CBG Case

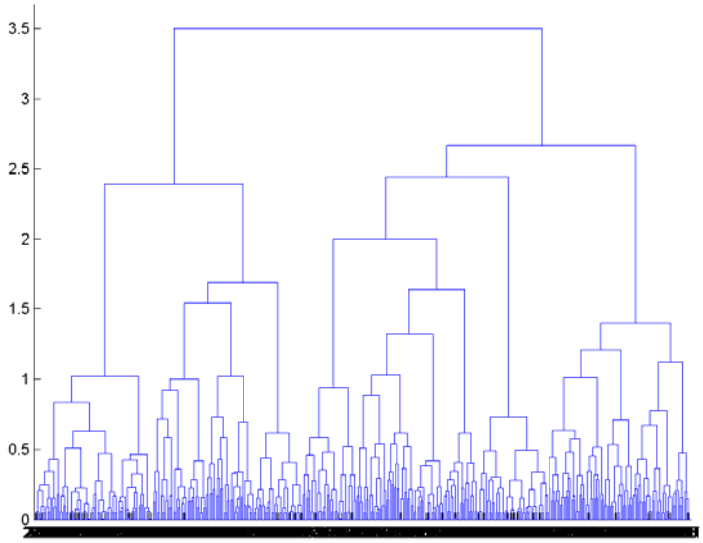


Figure 5.12 Dendrogram of Clustering Analysis with CDF + Categorical 2: CBG Case

The weighted mean squared errors (WMSE) computed are shown in Table 5.5. The highlighted entries indicate the lowest value for a given number of clusters. The entries with n.a. indicate that OLS model cannot be estimated due to singularity problem. School district ranking and deck dummy cause this issue since one generated cluster includes a single value for these two variables.

WMSE value indicates that categorical 2 method is highly likely the best choice for the census block group case. We estimated OLS by excluding school district ranking and/or deck variable to see the trend for WMSE for the cases we cannot calculate with all variables and confirmed the tendency. Therefore, we proceed with categorical 2 method for the census block group case.

N. Clusters	CDF	Hamming	Categorical 1	Categorical 2
1	600.22	600.22	600.22	600.22
2	358.80	292.43	315.17	317.70
3	277.56	221.09	215.75	187.96
4	257.76	164.45	159.84	157.98
5	254.44	156.87	129.32	129.62
6	187.39	150.21	120.56	112.34
7	128.64	134.07	115.64	106.12
8	124.13	84.65	109.36	103.11
9	73.26	82.94	101.25	94.32
10	71.26	73.12	67.88	67.55
11	n.a.	64.29	66.32	n.a.
12	n.a.	53.49	62.30	n.a.
13	n.a.	52.60	56.30	n.a.
14	n.a.	46.90	55.65	n.a.
15	n.a.	45.77	53.47	n.a.
16	n.a.	39.86	n.a.	n.a.
17	n.a.	39.25	n.a.	n.a.
18	n.a.	38.94	n.a.	n.a.
19	n.a.	n.a.	n.a.	n.a.
20	n.a.	n.a.	n.a.	n.a.

Table 5.5. WMSE Comparison Census Block Group Case

WMSE is plotted in Figure 5.13 together with WMSE computed from the results of OLS estimates without the school district ranking which causes singularity problem in order to see the trend after ten clusters. WMSEs computed from two specifications are very similar to each other. Therefore, WMSE plot without the school district ranking can be an appropriate alternative to find the “knee-point”.

WMSE falls until around six clusters, labels till nine clusters, drops again at ten clusters and becomes relatively flat again although the value is decreasing by small amount. Therefore, we consider cluster number ten could be a reasonable number of clusters.

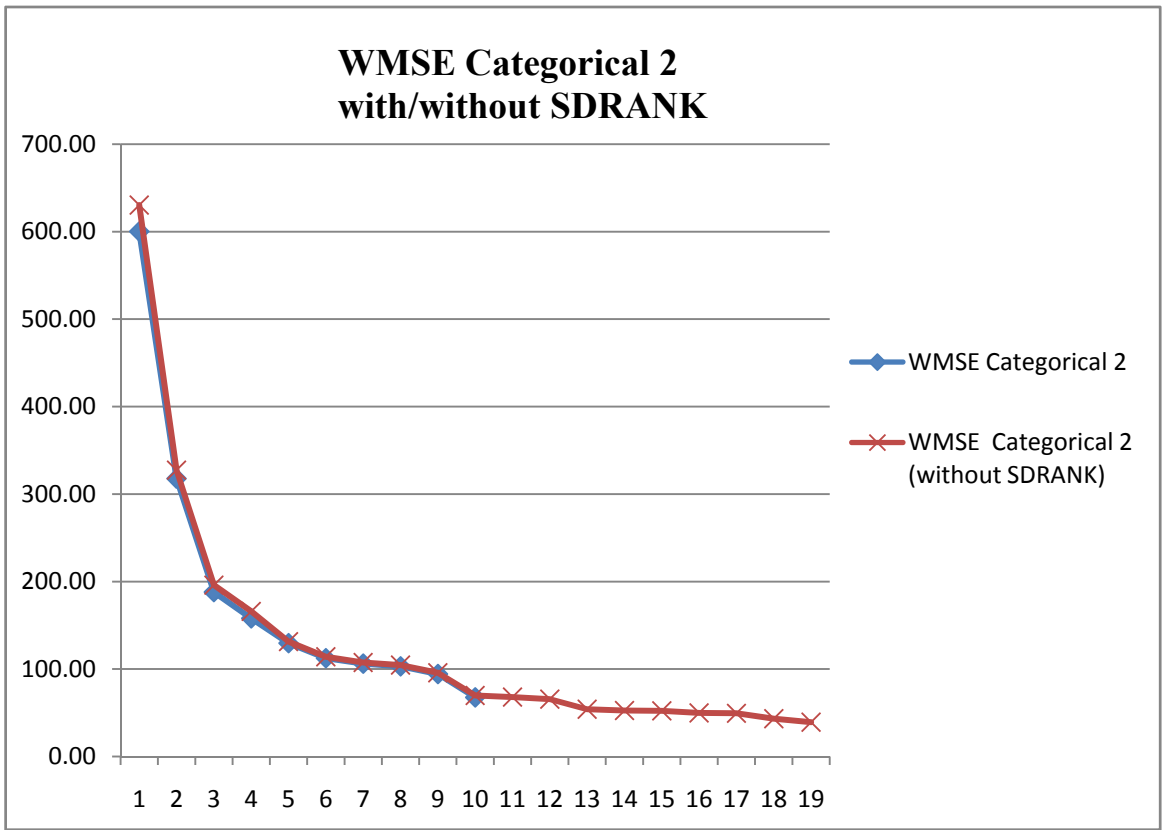


Figure 5.13. WMSE for CDF + Categorical 2 Clustering Method: CBG Case

	Categorical 2	Categorical 2 (without SDRANK)
1	0.697	0.682
2	0.692	0.682
3	0.682	0.672
4	0.684	0.674
5	0.672	0.669
6	0.662	0.659
7	0.653	0.650
8	0.655	0.651
9	0.652	0.648
10	0.662	0.654
11		0.652
12		0.653
13		0.655
14		0.656
15		0.658
16		0.659
17		0.660
18		0.657
19		0.657

Table 5.6. Calculated Weighted R-squares for Categorical 2 method: CBG Case

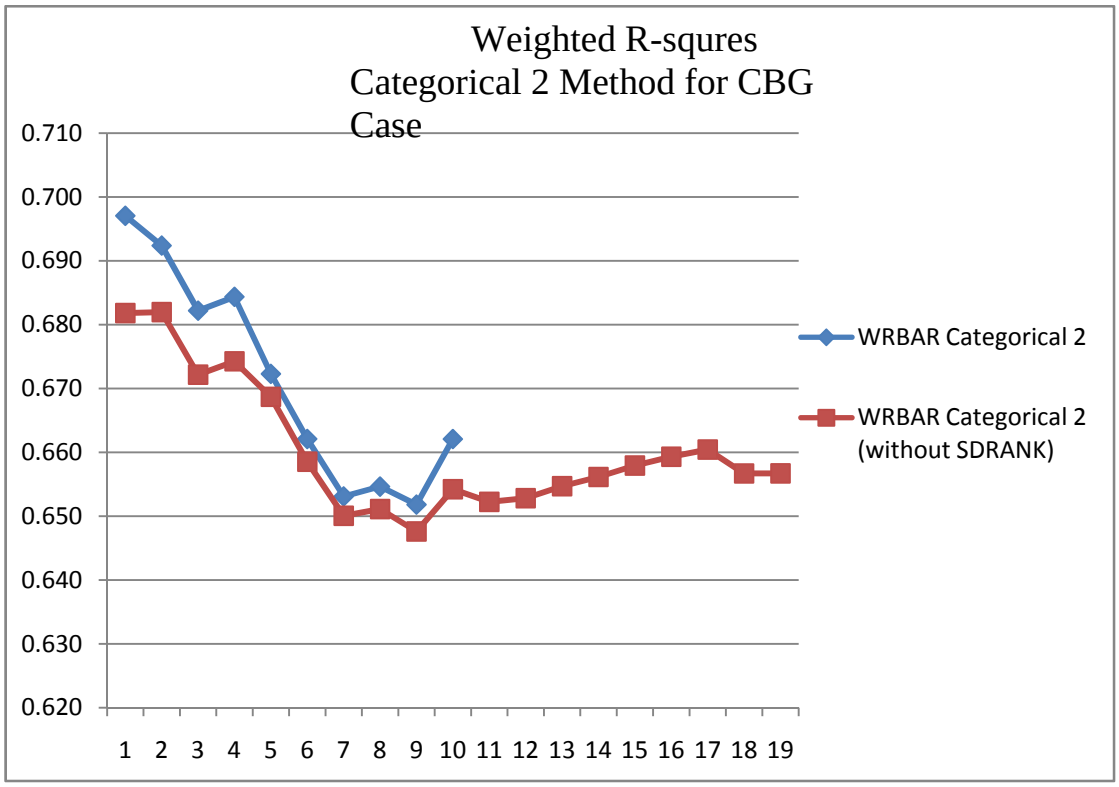


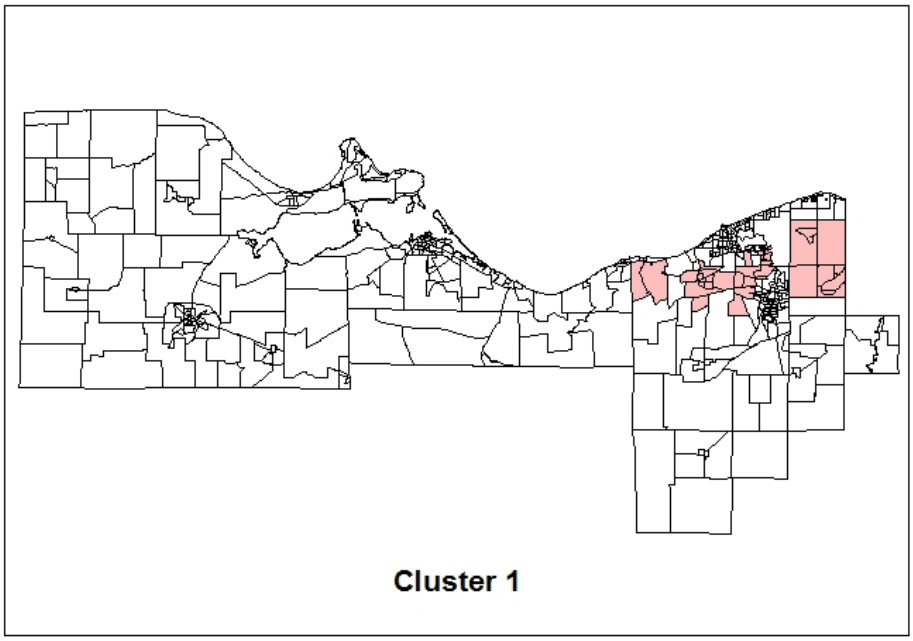
Figure 5.14. Plotted Weighted R-squares: CBG Case

Table 5.6 includes calculated weighted R-squares for the Categorical 2 method by using OLS with and without school district ranking and Figure 5.14 is the plot of both cases. We can observe two things from this figure. One is that although the case with school district ranking has higher weighted R-squares, the shapes of the figures look very similar to each other. The other is that after constantly decreasing until around seven clusters, it goes up again at ten clusters. If we assume that the trend from both specifications stays similar to each other, then we can expect the value to fall again after ten clusters and the next peak comes at 17 clusters.

Considering the analysis from both WMSE and Weighted R-squares, we decide to choose Categorical 2 method with ten clusters for census block group case, and proceed with our analysis with the specification.

5.3.2 Analysis of Clustering Outcomes

Clustered census block group locations are shown in Figure 5.15. If we compare the locations of the clusters from individual house case, the similarity between the two can be observed. Cluster 1+2 in the individual house case (IH) is similar to Cluster 8 of the CBG case. Similarly, Cluster 4 of IH and Cluster 1 of CBG, cluster 5 of IH and Cluster 9 of CBG, Cluster 7 of IH and Cluster 2 of CBG, Cluster 8 of IH and Cluster 6 of CBG, and Cluster 10 of IH to Cluster 10 of CBG are located in very similar places. One reason for the similarity could be the fact that we use census block group level median household income when we cluster individual houses. The other reason might be that the centroid becomes a good “average point” in terms of proximity measures. Therefore, although the clustering substances and dissimilarity measure used are different, we can conclude that outcome of the clustering with given set of filtering variables returns very similar set of clusters.



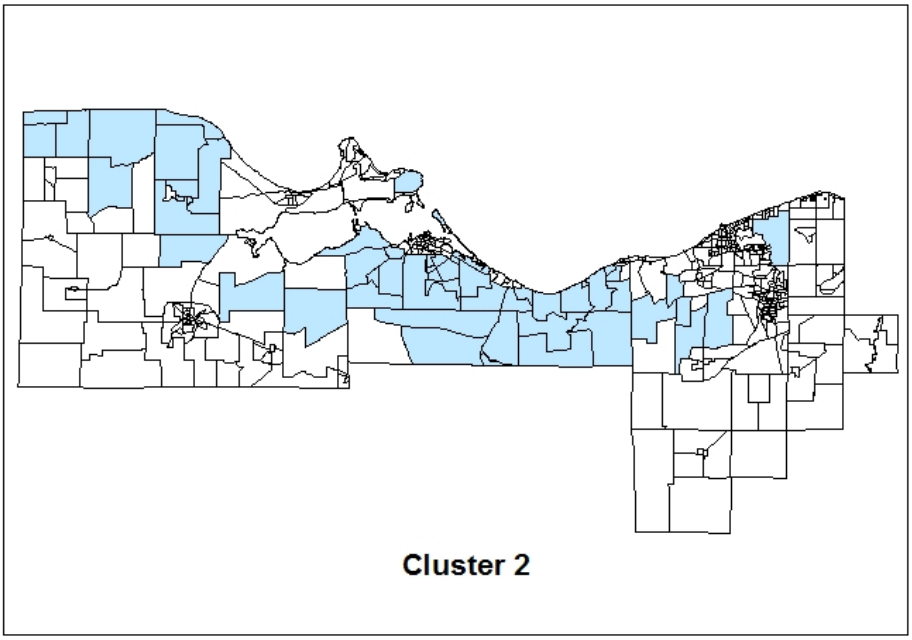
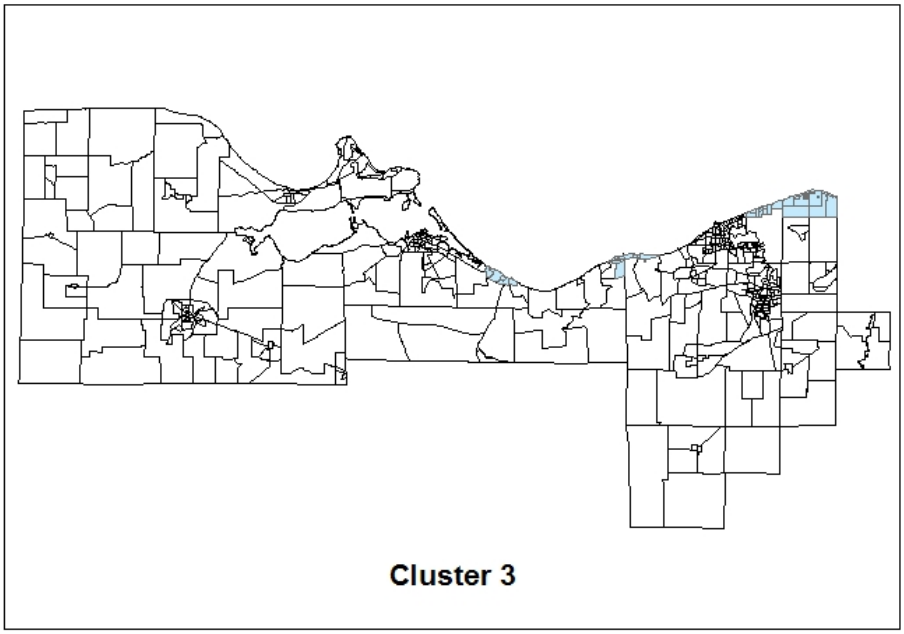


Figure 5.15 Census Block Groups in Each Cluster

Figure 5.15 (continued)



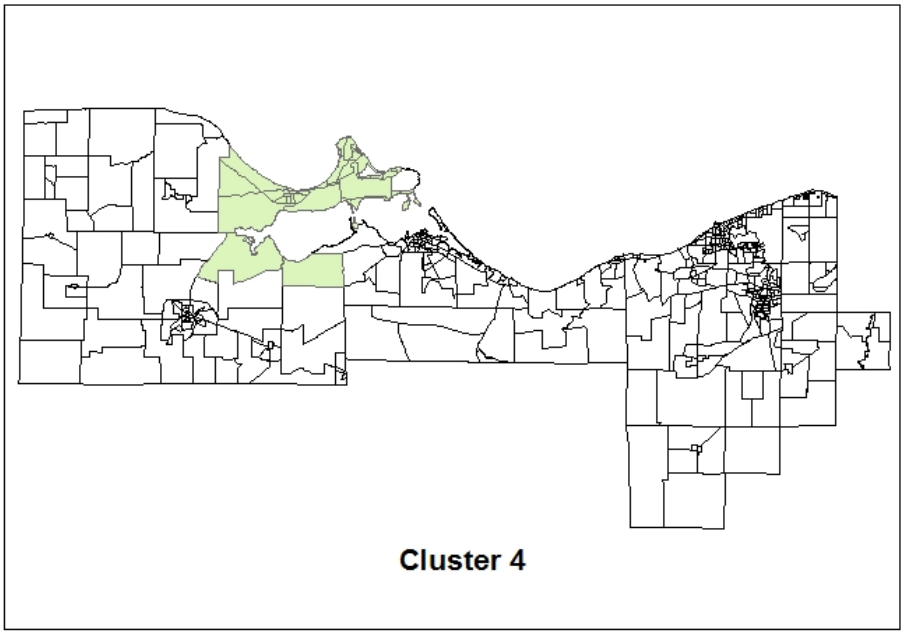
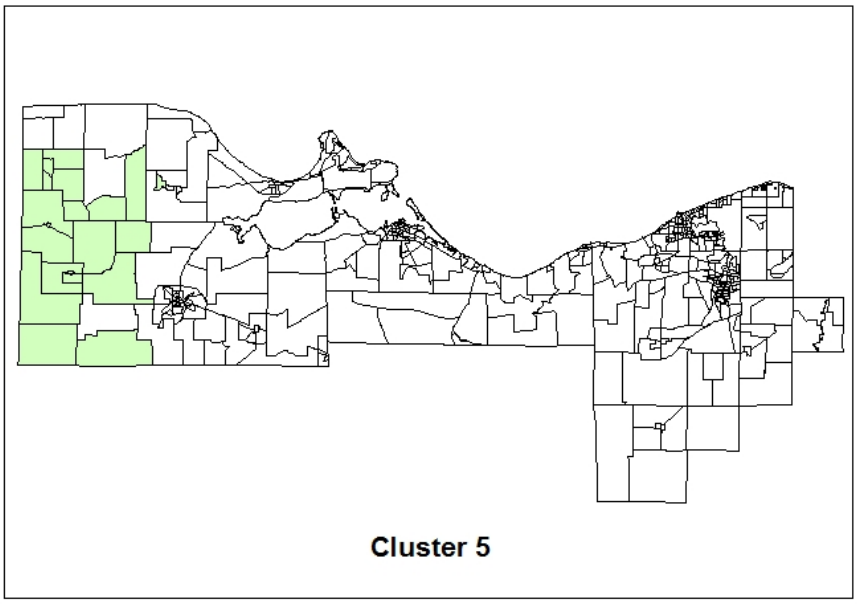


Figure 5.15(continue)

Figure 5.15 (continued)



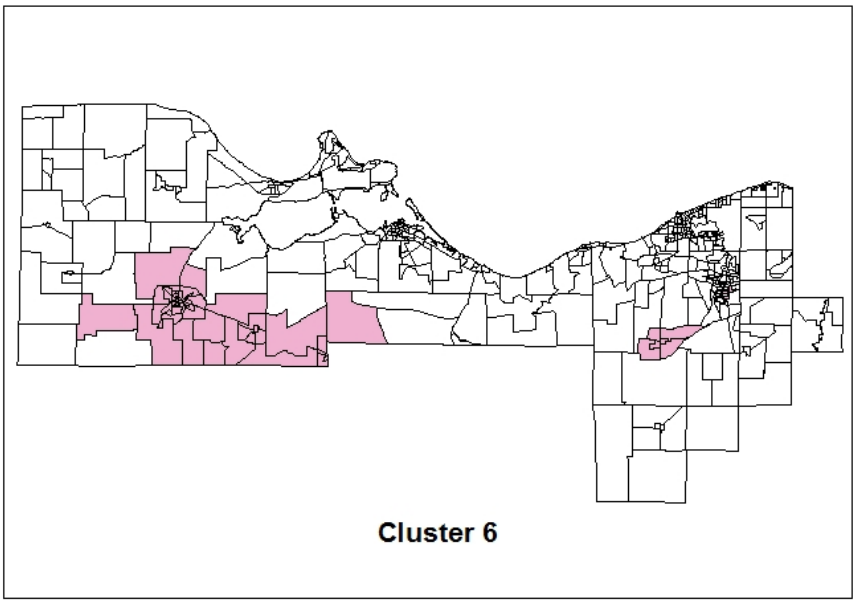
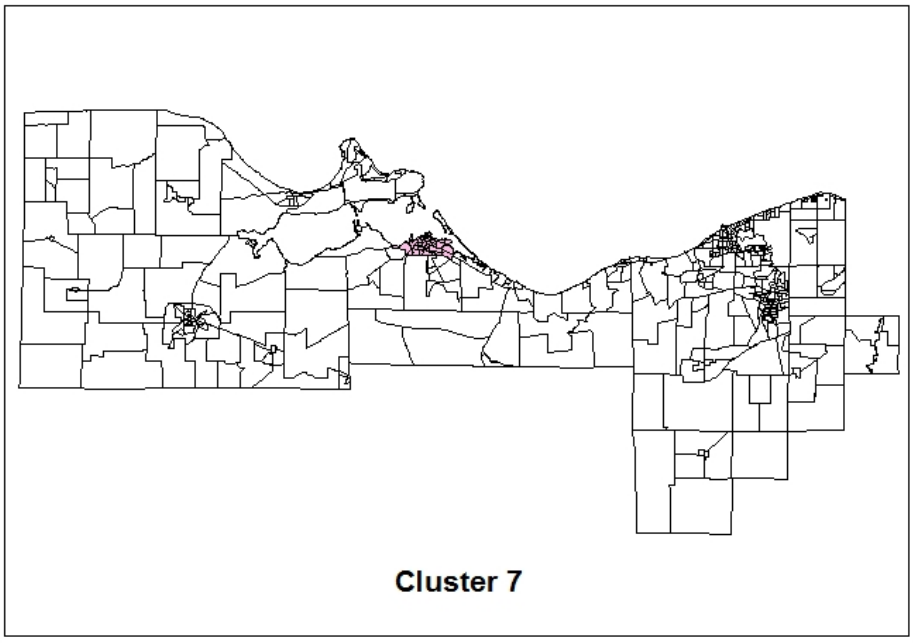


Figure 5.15 (continue)

Figure 5.15 (continued)



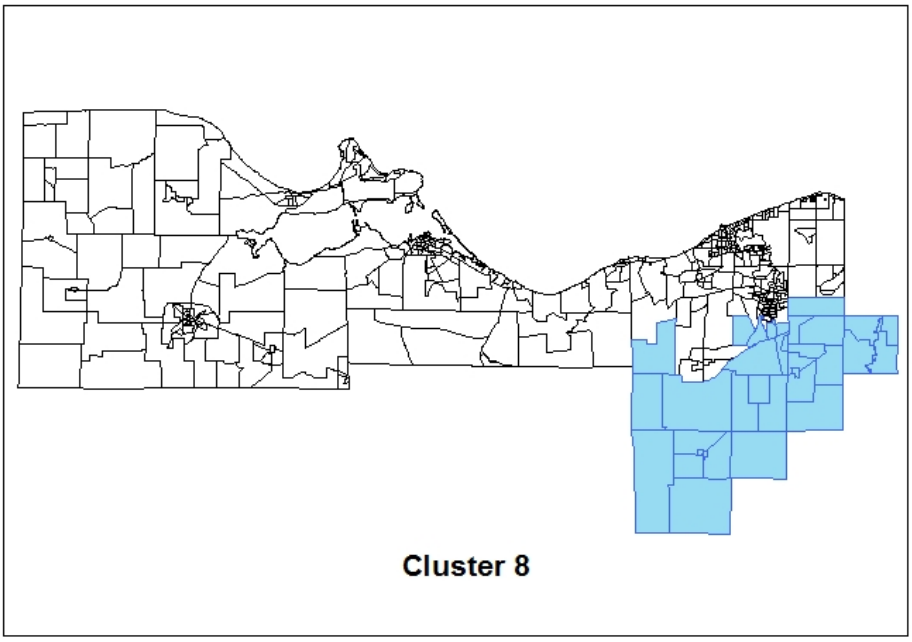
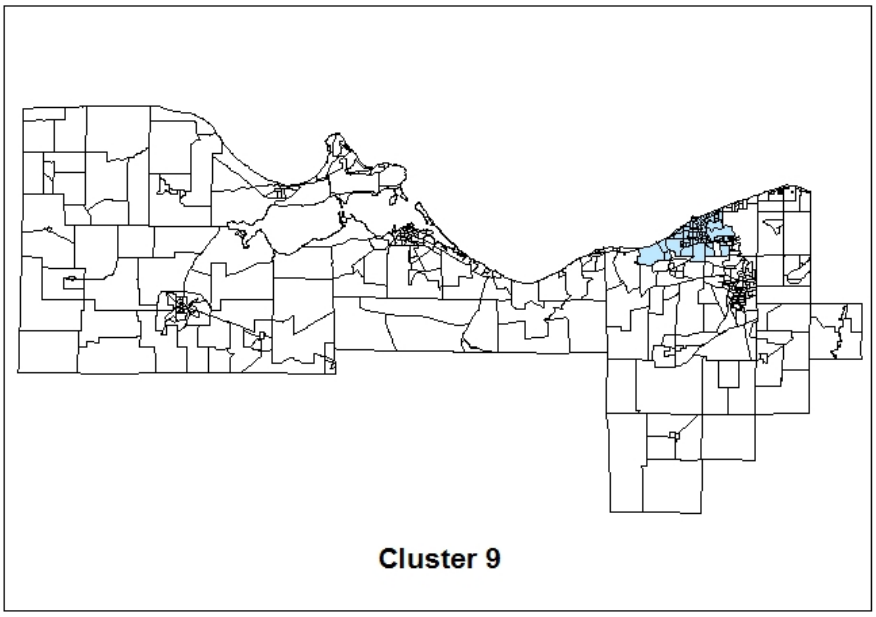
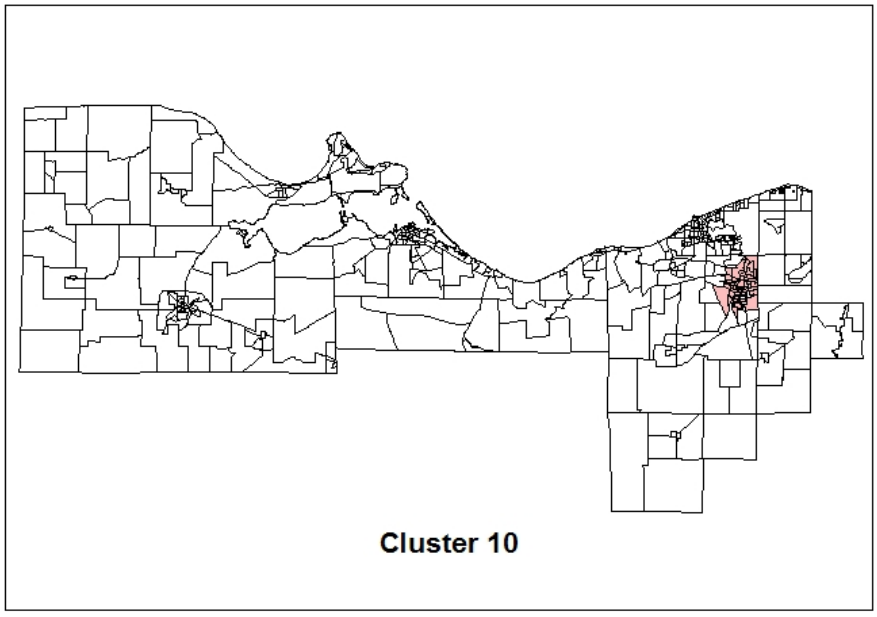


Figure 5.15 (continue)

Figure 5.15 (continued)





The descriptive statistics for each cluster are listed in Table 5.7. There are two most distinguishable clusters. Cluster 3 has the highest median household income value, the highest housing price, the largest building square feet, the newest in terms of the age of the house and the highest school district ranking. It has very short average distances to the closest city and to the coast line. Fecal coliform count is the worst for this cluster and water clarity marks the highest value. Cluster 7 has the opposite characteristics of Cluster 3 although they are both located very near to the lake. It has the lowest median household income, the lowest housing price, and the lowest average school district ranking. The average fecal coliform value is the lowest for this cluster and the water clarity is the lowest. The oldest houses are in Cluster 5 and it also has the highest average fecal coliform counts value.

	Cluster 1				Cluster 2			
	Mean	St.Dev.	Min.	Max	Mean	St.Dev.	Min.	Max
MEDHHINC	52239	8202	31718	62852	51946	8226	31345	72807
CITY	5.27	2.17	0.85	8.34	9.89	5.42	4.02	28.02
COAST	7.92	2.85	3.15	13.15	4.82	3.66	0.49	12.98
DPRICE	117391	46334	50078	379900	112551	57326	50000	582000
LOTACR	500.50	1093.66	20.00	19180.00	1027.40	4077.59	41.00	78000.00
BLDGSF	1716.87	528.85	538.00	4796.00	1606.23	617.00	204.00	4868.00
BATHN	1.51	0.55	1.00	4.00	1.32	0.52	1.00	3.00
GRGSQF	38.75	135.55	0.00	1200.00	410.63	279.02	0.00	3651.00
AGE	22.44	21.09	0.00	171.00	32.64	24.63	1.00	171.00
AIRCNDD	0.94	0.24	0.00	1.00	0.35	0.48	0.00	1.00
DECKD	0.08	0.27	0.00	1.00	0.19	0.39	0.00	1.00
FIREPLD	0.55	0.50	0.00	1.00	0.40	0.49	0.00	1.00
SDRANK	13.73	9.73	1.00	38.00	13.98	9.47	2.00	37.00
FECAL	266.86	228.53	29.28	2717.26	216.97	412.19	12.00	2717.26
SECCHI	237.72	75.52	128.96	431.78	186.73	52.10	89.54	355.58
N	2628				781			

	Cluster 3				Cluster 4			
	Mean	St.Dev.	Min.	Max	Mean	St.Dev.	Min.	Max
MEDHHINC	62045	11700	30340	75905	43000	8757	21363	60462
CITY	2.97	0.97	0.51	5.99	8.22	6.27	0.82	17.80
COAST	1.43	0.81	0.15	2.65	1.01	0.72	0.01	3.39
DPRICE	165824	94381	50427	669292	100568	53645	50000	372250
LOTACR	359.01	452.30	16.00	10000.00	316.44	472.15	21.00	3420.00
BLDGSF	1990.63	810.88	196.00	5824.00	1367.87	475.22	504.00	3874.00
BATHN	1.66	0.62	1.00	5.00	1.35	0.56	1.00	4.00
GRGSQF	109.62	200.74	0.00	912.00	303.34	267.40	0.00	1140.00
AGE	21.79	20.92	0.00	164.00	45.31	28.71	1.00	121.00
AIRCNDD	0.79	0.41	0.00	1.00	0.21	0.41	0.00	1.00
DECKD	0.13	0.34	0.00	1.00	0.20	0.40	0.00	1.00
FIREPLD	0.68	0.47	0.00	1.00	0.33	0.47	0.00	1.00
SDRANK	6.89	5.64	2.00	38.00	16.81	4.30	13.00	22.00
FECAL	369.47	321.89	18.24	2717.26	129.82	339.00	12.00	2528.42
SECCHI	239.94	78.88	89.54	431.78	203.07	40.89	118.96	258.80
	1494				280			

Table 5.7. Descriptive Statistics for Each Cluster: CBG Case

Table 5.7 (continued)

	Cluster 5				Cluster 6			
	Mean	St.Dev.	Min.	Max	Mean	St.Dev.	Min.	Max
MEDHHINC	41935	5498	31470	55093	41014	7845	19703	57099
CITY	22.43	3.31	11.85	27.23	2.98	1.93	0.55	13.91
COAST	20.29	4.23	11.22	30.35	14.26	2.21	8.83	23.71
DPRICE	85847	28911	50000	200558	86189	37321	50000	506000
LOTACR	842.92	4102.40	56.00	72000.00	394.15	674.81	10.00	10690.00
BLDGSF	1535.84	480.80	660.00	3750.00	1523.52	518.60	202.00	4592.00
BATHN	1.23	0.43	1.00	3.00	1.20	0.43	1.00	5.00
GRGSQF	406.70	264.63	0.00	1440.00	311.99	261.10	0.00	1200.00
AGE	54.78	31.41	3.00	162.00	51.39	29.56	0.00	159.00
AIRCNDD	0.21	0.41	0.00	1.00	0.39	0.49	0.00	1.00
DECKD	0.10	0.30	0.00	1.00	0.11	0.31	0.00	1.00
FIREPLD	0.31	0.46	0.00	1.00	0.28	0.45	0.00	1.00
SDRANK	16.70	8.26	10.00	31.00	23.26	5.88	4.00	33.00
FECAL	515.21	571.88	14.15	1709.34	109.24	127.09	14.15	869.98
SECCHI	175.22	45.04	116.93	242.43	185.80	58.46	119.23	398.75
	323				1118			

	Cluster 7				Cluster 8			
	Mean	St.Dev.	Min.	Max	Mean	St.Dev.	Min.	Max
MEDHHINC	33050	5817	9113	47221	50413	7088	31963	60618
CITY	2.87	1.34	0.50	5.47	11.12	4.22	5.70	23.61
COAST	1.15	0.62	0.03	2.43	21.32	4.81	13.68	36.87
DPRICE	79038	37107	50000	385000	113343	47097	50000	465841
LOTACR	363.41	2707.62	50.00	43000.00	1630.57	2559.94	75.00	18780.00
BLDGSF	1496.67	558.51	204.00	4032.00	1680.34	606.58	480.00	5506.00
BATHN	1.12	0.35	1.00	3.00	1.48	0.58	1.00	4.00
GRGSQF	347.51	219.44	0.00	1232.00	64.37	205.63	0.00	4040.00
AGE	53.06	26.96	1.00	155.00	22.90	20.27	0.00	150.00
AIRCNDD	0.27	0.45	0.00	1.00	0.90	0.30	0.00	1.00
DECKD	0.06	0.24	0.00	1.00	0.08	0.28	0.00	1.00
FIREPLD	0.22	0.41	0.00	1.00	0.51	0.50	0.00	1.00
SDRANK	36.64	3.27	7.00	37.00	19.33	8.04	1.00	36.00
FECAL	65.66	25.79	46.20	142.86	227.88	209.05	29.28	2717.26
SECCHI	156.95	37.06	92.60	195.71	234.75	74.42	128.96	431.78
	251				1284			

Table 5.7 (continue)

Table 5.7 (continued)

	Cluster 9				Cluster 10			
	Mean	St.Dev.	Min.	Max	Mean	St.Dev.	Min.	Max
MEDHHINC	40489	10852	12902	74166	42466	10221	11451	69082
CITY	3.43	1.51	0.52	6.87	4.45	1.56	0.20	6.45
COAST	2.30	1.64	0.14	6.23	12.31	2.07	9.80	15.43
DPRICE	88371	36404	50000	311393	92682	35341	50000	334466
LOTACR	251.48	251.35	46.00	3920.00	241.86	208.69	10.00	3730.00
BLDGSF	1490.21	488.38	640.00	4699.00	1493.15	514.56	576.00	5074.00
BATHN	1.29	0.51	1.00	4.00	1.35	0.54	1.00	5.00
GRGSQF	20.10	91.74	0.00	766.00	56.24	152.38	0.00	835.00
AGE	32.86	17.71	0.00	94.00	28.58	19.07	0.00	128.00
AIRCNDD	0.96	0.20	0.00	1.00	0.90	0.31	0.00	1.00
DECKD	0.08	0.27	0.00	1.00	0.08	0.27	0.00	1.00
FIREPLD	0.38	0.48	0.00	1.00	0.41	0.49	0.00	1.00
SDRANK	33.09	11.21	6.00	38.00	31.58	3.97	4.00	33.00
FECAL	201.85	154.45	29.28	869.98	323.14	257.01	29.28	869.98
SECCHI	223.82	61.01	147.07	398.75	229.51	73.38	147.07	398.75
	1124				1382			

5.4 Conclusion

Hierarchical clustering with four different similarity measures are implemented and compared. We selected one measure by comparing calculated weighted mean squared error (WMSE) and determined the number of clusters by finding the “knee-point” and also looking at weighted R-squares computed. For the individual houses case, we selected the Categorical 1 method with 11 number of clusters while the Categorical 2 method with 10 number of clusters was chosen for the census block group case.

As the comparison between the clusters generated by using two different data sets (individual houses and census block groups), we found that six out of ten clusters are

located in very similar places. Therefore, we can conclude that for the similar clusters, the centroids of the census block groups played a role as a reasonable representative of the houses located within the census block groups, and the generated clusters are relatively robust to the types of data and types of similarity measure we adopted.

CHAPTER 6

THE FIRST STAGE OF THE HEDONIC MODEL ON LAKE ERIE WATER QUALITY

6.1 The Model

6.1.1 Ordinary Least Squares (OLS)

Since the hedonic price function is the locus of equilibrium points, there is little a priori information to determine the functional form. Following the findings from Cropper, Deck and McConnell (1988), four simple functional forms such as linear, double-log, semi-log and inverse semi-log are our possible choices.

Linear: $$P = \alpha + \sum_i \beta_i H_i + \sum_j \gamma_j N_j + \sum_k \mu_k D_k + \sum_l \theta_l E_l + \varepsilon$$

Semi-log: $$\ln P = \alpha + \sum_i \beta_i H_i + \sum_j \gamma_j N_j + \sum_k \mu_k D_k + \sum_l \theta_l E_l + \varepsilon$$

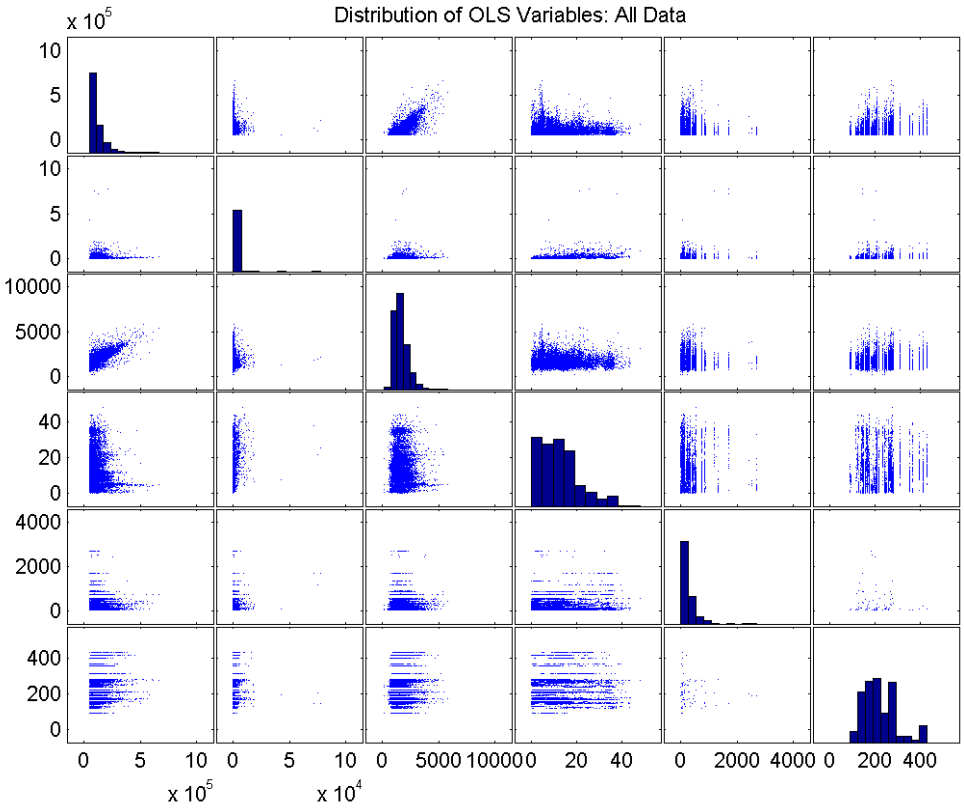
Log-linear: $$P = \alpha + \sum_i \beta_i \ln H_i + \sum_j \gamma_j \ln N_j + \sum_k \mu_k \ln D_k + \sum_l \theta_l \ln E_l + \varepsilon$$

Log-log: $$\ln P = \alpha + \sum_i \beta_i \ln H_i + \sum_j \gamma_j \ln N_j + \sum_k \mu_k \ln D_k + \sum_l \theta_l \ln E_l + \varepsilon$$

By examining the distribution of each variable (diagonal figures in Figure 6.1. for all data, Figure 6.2 – 6.12 for each cluster), we choose to take the logarithm of the following variables.

- Discounted Housing Price (base year = 1996)
- Lot acreage
- Building Square Feet
- Distance to the Closest Beach
- Fecal
- Secchi

Since the Garage Square Feet variable contains sufficiently large number of zeros indicating there is no garage at the house, we decide not to take the logarithm of this variable. In order to incorporate the vintage effect, we include the squared house age variable in addition to the linear age variable. Except for school district ranking, all the variables are significant.



*From left to right (and top to bottom), variables are DPRICE, LOTACR, BLDGSF, BEACHDIST, FECAL and SECCHI. Diagonal figures are the histograms of individual variables, and off-diagonal figures show the columns of X plotted against the columns of Y.

Figure 6.1. Distribution of OLS Variables: All Data

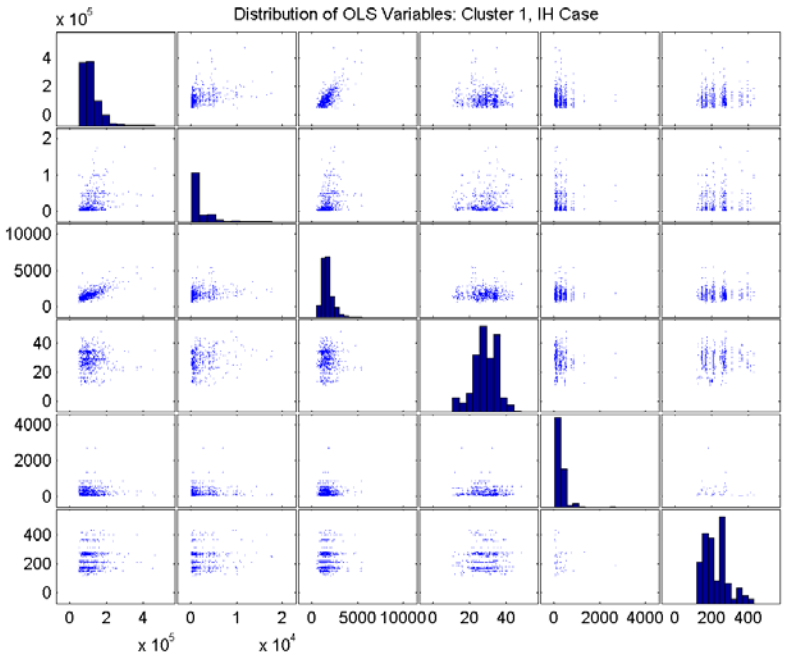


Figure 6.2. Distribution of OLS Variables: Cluster 1, IH Case

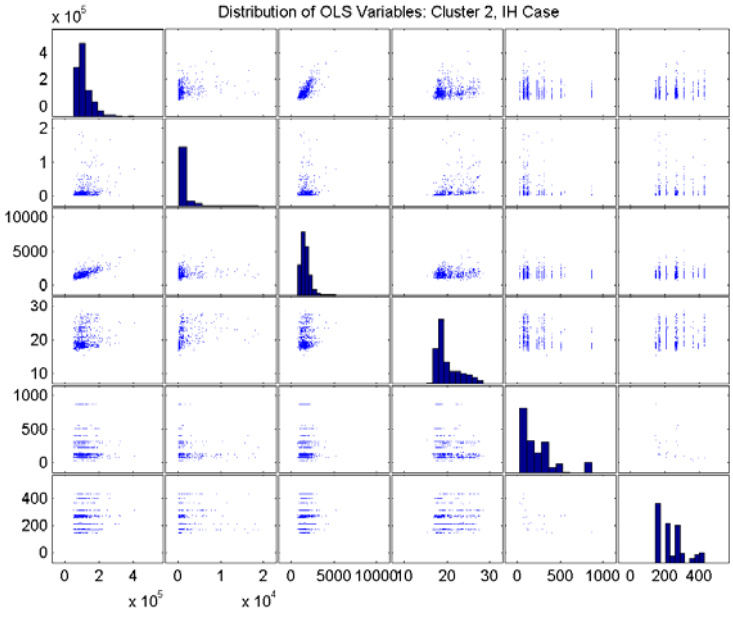


Figure 6.3. Distribution of OLS Variables: Cluster 2, IH Case

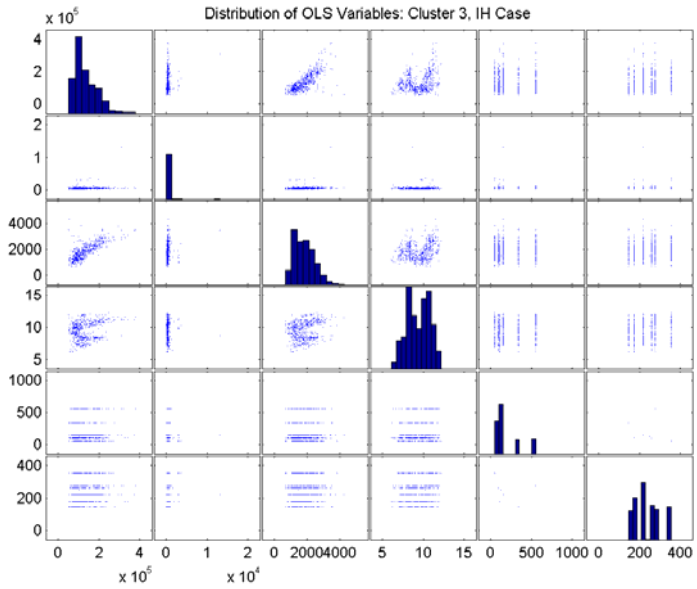


Figure 6.4. Distribution of OLS Variables: Cluster 3, IH Case

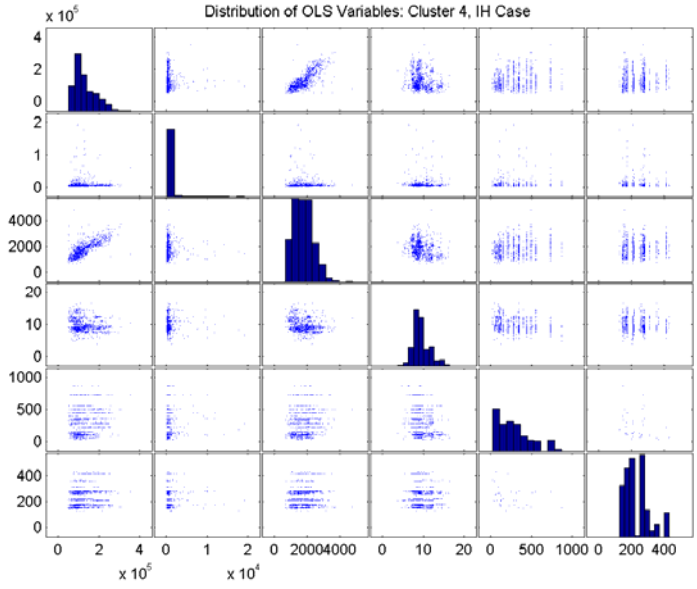


Figure 6.5. Distribution of OLS Variables: Cluster 4, IH Case

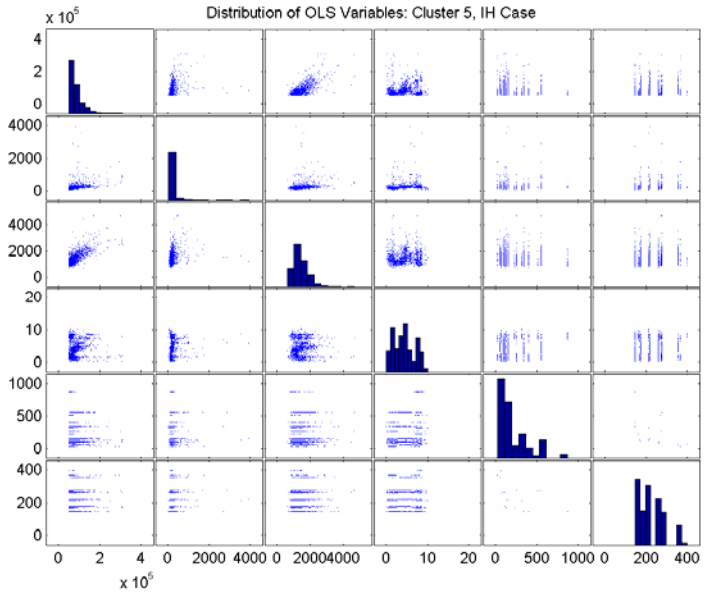


Figure 6.6. Distribution of OLS Variables: Cluster 5, IH Case

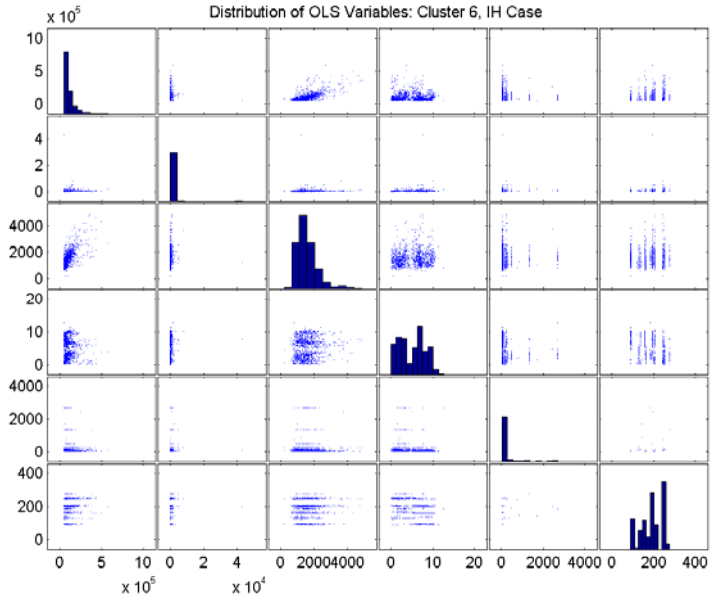


Figure 6.7. Distribution of OLS Variables: Cluster 6, IH Case

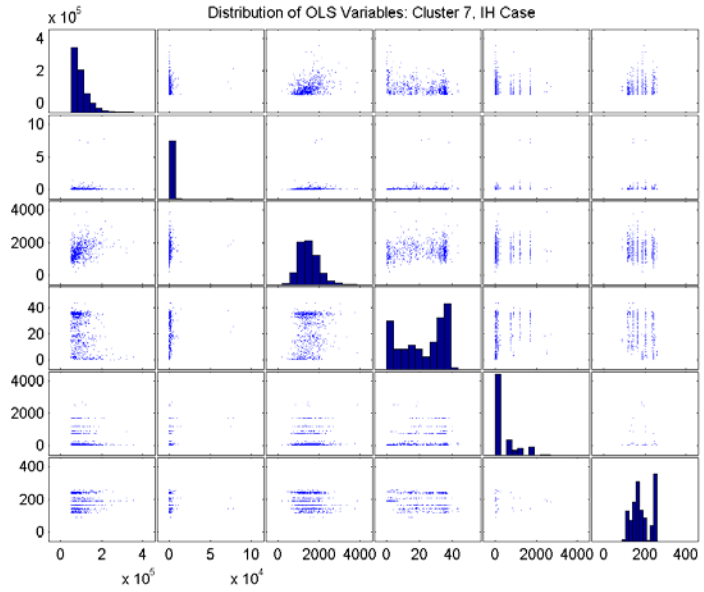


Figure 6.8. Distribution of OLS Variables: Cluster 7, IH Case

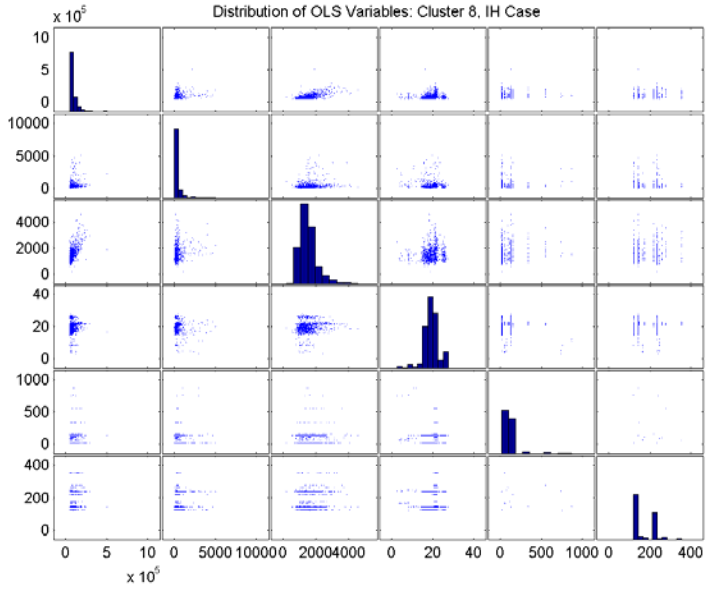


Figure 6.9. Distribution of OLS Variables: Cluster 8, IH Case

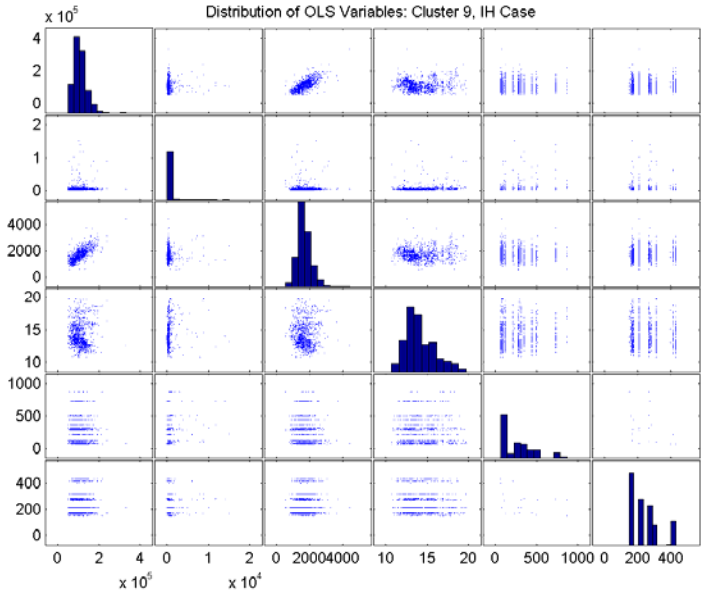


Figure 6.10. Distribution of OLS Variables: Cluster 9, IH Case

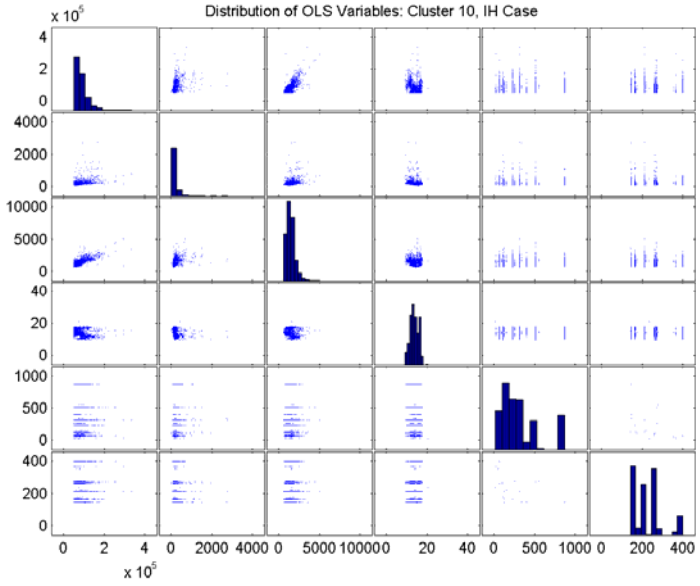


Figure 6.11. Distribution of OLS Variables: Cluster 10, IH Case

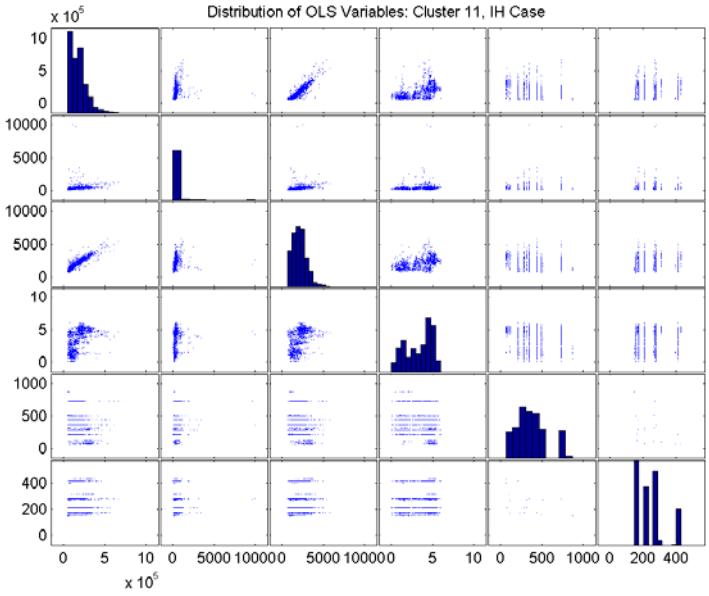


Figure 6.12. Distribution of OLS Variables: Cluster 11, IH Case

6.1.2 Spatial Hedonic Model

GMM methodology introduced by Kelejian and Prucha (1998) is used for the estimation of the spatial hedonic price function. In order to determine the spatial hedonic model specification and the appropriate weight matrix, we tested weight matrices with four different cutoff distances, 200, 400, 800 and 1600 by using robust Lagrange Multiplier (LM) test (See Section 2.3.1) for both spatial lag and error specifications. We first determine whether spatial lag or error model is the likely model by examining the robust LM test outcomes. Given the knowledge of the spatial model specification, we choose the weight matrix out of four types according to the level of goodness of model fit.

Both spatial lag and error model specifications used for the test are expressed as follows. We found that in our analysis, all of the best fitted models have spatial error specification.

Spatial Lag Model Specification

$$\begin{aligned}\ln P = \alpha + \rho W \ln P + \beta_1 \ln LOTACR + \beta_2 \ln BLDGSF + \beta_3 BATH + \beta_4 GRGSQF + \beta_5 AGE \\ + \beta_6 AGE^2 + \beta_7 AIRCNDD + \beta_8 DECKD + \beta_9 FIREPLD + \beta_{10} SDRANK \\ + \beta_{11} \ln BEACH + \beta_{12} \ln FECAL + \beta_{13} \ln SECCHI + \varepsilon\end{aligned}$$

Spatial Error Model Specification

$$\begin{aligned}\ln P = \alpha + \beta_1 \ln LOTACR + \beta_2 \ln BLDGSF + \beta_3 BATH + \beta_4 \ln GRGSF + \beta_5 AGE \\ + \beta_6 AGE^2 + \beta_7 AIRCNDD + \beta_8 DECKD + \beta_9 FIREPLD + \beta_{10} SDRANK \\ + \beta_{11} \ln BEACH + \beta_{12} \ln FECAL + \beta_{13} \ln SECCHI + \lambda W\varepsilon + u\end{aligned}$$

6.2 Estimated Results: Individual Houses Case

In this section, we report the estimated results of estimated OLS results, corresponding to the Chow test, the robust LM test results, the estimated GMM results and the derived marginal implicit prices (MIP) for each cluster for the individual houses case.

6.2.1 Estimated Result of OLS

The estimated results of OLS are listed in Table 6.1. Lot acreage and building square feet are significant for all clusters at least at 10 percent level with the expected sign. Bathroom variable is statistically significant for all clusters except for cluster 3. On the other hand, garage square feet is not statistically significant for most of the clusters

except for cluster 6, 7 and 8. It is interesting to notice that the mean values of this value are more than 300 square feet while the overall mean is 133 square feet. This result indicates that for these clusters, having large garage is an important factor. Age of the house is negative significant for all but cluster 10 and vintage effect (Age squared) is positive significant for all clusters but cluster 6 and 10. Air-conditioning and fireplace dummy variables are positively significant for all clusters while deck dummy is not significant for four clusters, 4, 5, 7, and 8. School district ranking is negative significant for six clusters.

Distance to the closest beach has mixed output. It is negative significant meaning being closer to the beach is valued in clusters 1, 5, 6, 7, 9 and 10 while it is positive significant in clusters 3, 8 and 11. Fecal coliform counts variable also has the mixed results. The expected sign of the variable is negative. We found it negative significant for clusters 6, 9 and 11 while positive significant for clusters 4, 7, 8 and 10.

We analyzed the results from different clustering methods as well as different specification of the models apart from the setting reported here, these positive significant fecal variables persist. The expected sign of the secchi variable is positive, and it is positively significant for clusters of 1, 2, 4, 6, 8 and 10. Positively significant result for the secchi variable is consistent across different clustering methods and model specifications.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
CONSTANT	7.784	***	7.231	***	7.600	***	6.620	***	7.508	***
	(22.16)		(19.55)		(17.91)		(19.77)		(32.32)	
LNLOTACR	0.072	***	0.083	***	0.038	**	0.066	***	0.163	***
	(8.93)		(8.92)		(1.93)		(5.67)		(11.83)	
LNBLDGSF	0.437	***	0.445	***	0.445	***	0.466	***	0.400	***
	(12.56)		(12.93)		(10.18)		(12.43)		(15.40)	
BATHN	0.048	**	0.064	***	0.007		0.068	***	0.102	***
	(2.42)		(3.74)		(0.30)		(3.52)		(7.20)	
GRGSQF	-0.00002		-0.0001		0.0001		-0.00001		0.0001	
	(-0.40)		(-1.26)		(0.40)		(-0.07)		(1.33)	
AGE	-0.006	***	-0.011	***	-0.010	***	-0.007	***	-0.010	***
	(-6.15)		(-9.38)		(-8.27)		(-8.11)		(-8.97)	
AGE2	0.00004	***	0.0001	***	0.0001	***	0.00004	***	0.0001	***
	(4.46)		(5.78)		(5.44)		(5.87)		(6.64)	
AIRCNDD	0.094	**	0.141	***	0.189	*	0.191	***	0.178	***
	(2.54)		(3.62)		(1.79)		(3.97)		(4.13)	
DECKD	0.126	***	0.101	***	0.063	**	0.048		0.020	
	(4.37)		(3.42)		(2.21)		(1.58)		(0.91)	
FIREPLD	0.083	***	0.101	***	0.132	***	0.085	***	0.086	***
	(4.09)		(6.14)		(5.42)		(4.55)		(6.13)	
SDRANK	-0.001		-0.001		-0.004		-0.003	***	-0.002	***
	(-0.58)		(-0.75)		(-1.33)		(-2.93)		(-3.77)	
LNBEACH	-0.094	**	0.027		0.140	**	0.072		-0.033	***
	(-2.32)		(0.32)		(2.08)		(1.58)		(-3.75)	
LNFECAL	-0.009		0.011		-0.0002		0.040	***	-0.005	
	(-0.70)		(0.95)		(-0.02)		(3.50)		(-0.59)	
LNSECCHI	0.065	*	0.054	*	0.038		0.117	***	0.009	
	(1.83)		(1.87)		(0.98)		(4.24)		(0.40)	
AdjR2	0.57		0.64		0.66		0.67		0.66	
N	718		782		544		839		1229	

Table 6.1. Estimated Result of OLS for Each Cluster: Individual Houses Case

Table 6.1 (continued)

	Cluster 6		Cluster 7		Cluster 8		Cluster 9		Cluster 10	
CONSTANT	6.533	***	9.215	***	5.766	***	8.292	***	6.737	***
	(26.33)		(25.56)		(21.61)		(34.78)		(25.17)	
LNLOTACR	0.113	***	0.020	*	0.057	***	0.073	***	0.102	***
	(8.62)		(1.95)		(6.03)		(10.13)		(8.15)	
LNBLDGSF	0.480	***	0.317	***	0.514	***	0.359	***	0.448	***
	(16.09)		(8.50)		(19.97)		(13.45)		(20.67)	
BATHN	0.132	***	0.095	***	0.065	***	0.083	***	0.067	***
	(6.63)		(3.67)		(3.47)		(7.42)		(5.53)	
GRGSQF	0.0002	***	0.0002	***	0.0001	***	0.0001		-0.00003	
	(4.50)		(5.92)		(4.10)		(1.09)		(-0.58)	
AGE	-0.005	***	-0.006	***	-0.005	***	-0.010	***	-0.001	
	(-4.71)		(-4.27)		(-5.53)		(-13.83)		(-1.22)	
AGE2	0.00001		0.00004	***	0.00002	**	0.0001	***	-0.000004	
	(1.51)		(3.88)		(2.33)		(11.39)		(-0.35)	
AIRCNDD	0.064	***	0.143	***	0.033	*	0.119	***	0.099	***
	(3.40)		(5.05)		(1.83)		(4.96)		(3.56)	
DECKD	0.077	***	0.033		0.032		0.046	**	0.082	***
	(3.33)		(1.10)		(1.50)		(2.42)		(4.08)	
FIREPLD	0.078	***	0.147	***	0.132	***	0.084	***	0.092	***
	(4.02)		(6.37)		(7.11)		(7.36)		(7.19)	
SDRANK	0.0004		0.003	*	-0.006	***	0.014	***	0.008	**
	(0.52)		(1.94)		(-2.99)		(5.93)		(2.49)	
LNBEACH	-0.068	***	-0.134	***	0.159	***	-0.074	*	-0.087	**
	(-6.38)		(-13.03)		(4.32)		(-1.80)		(-2.22)	
LNFECAL	-0.027	***	0.014	*	0.064	***	-0.017	***	0.024	***
	(-3.16)		(1.95)		(8.10)		(-2.59)		(2.96)	
LNSECCHI	0.172	***	-0.033		0.181	***	0.018		0.087	***
	(5.95)		(-0.73)		(6.22)		(1.18)		(3.89)	
AdjR2	0.60		0.49		0.65		0.66		0.65	
N	1152		693		971		1185		1334	

Table 6.1 (continue)

Table 6.1 (continued)

	Cluster 11		Cluster 1+2	
CONSTANT	6.790	***	7.652	***
	(25.52)		(32.75)	
LNLOTACR	0.131	***	0.076	***
	(10.06)		(13.88)	
LNBLDGSF	0.638	***	0.444	***
	(21.45)		(18.41)	
BATHN	0.085	***	0.055	***
	(5.46)		(4.25)	
GRGSQF	-0.00001		-0.00003	
	(-0.27)		(-0.84)	
AGE	-0.008	***	-0.008	***
	(-7.89)		(-12.09)	
AGE2	0.0001	***	0.00005	***
	(4.23)		(8.08)	
AIRCNDD	0.137	***	0.122	***
	(4.75)		(4.89)	
DECKD	0.086	***	0.116	***
	(4.60)		(5.67)	
FIREPLD	0.064	***	0.094	***
	(3.64)		(7.29)	
SDRANK	-0.010	***	-0.001	
	(-6.83)		(-0.95)	
LNBEACH	0.020	**	-0.092	***
	(2.03)		(-3.80)	
LNFECAL	-0.074	***	0.000	
	(-6.29)		(0.04)	
LNSECCHI	-0.033		0.060	***
	(-1.60)		(2.68)	
Rbar-squared	0.85		0.60	
N	1218		1500	

Given the OLS outcomes, the Chow test is implemented to see which clusters actually have statistically distinguishable coefficients from others. We compared each cluster one by one for all possible combinations. The results can be found in Table 6.2. The F value

threshold for our case is 2.039 for one percent significance level. Since clusters 1 and 2 have the F values less than the threshold value, we merged these two clusters to form one market. Chow test result of the merged cluster against all others is listed in the bottom of the table. Since all other clusters have F values greater than the threshold value, we proceed with ten clusters (1+2, 3 – 11) in the following analysis.

	Cluster										
	1	2	3	4	5	6	7	8	9	10	11
1	-	1.89	2.50	2.47	5.73	4.32	4.59	7.12	4.52	3.56	22.29
2		-	3.43	1.96	4.43	5.41	7.24	8.71	4.06	4.69	20.04
3			-	2.33	8.39	5.27	3.93	3.59	9.04	6.63	15.99
4				-	6.66	5.48	7.02	3.72	7.14	4.75	26.81
5					-	9.89	23.60	12.09	11.02	8.77	22.99
6						-	12.83	7.18	13.73	9.23	17.82
7							-	10.06	8.39	12.95	36.68
8								-	15.45	11.38	17.33
9									-	9.90	34.82
10										-	21.24
11											-
1+2	-	-	2.94	2.20	7.55	6.84	7.74	9.64	4.53	4.74	30.88

Table 6.2. Chow Test Result: Individual Houses Case

6.2.2 Estimated Result of Spatial Hedonic Model

Given the OLS outcomes and four types of weight matrices (threshold values of 200, 400, 800 and 1600 meters), the robust LM test is implemented to specify the likely spatial model and the appropriate weight matrix for each cluster. Weight matrices are generated

by using the GEODA application (downloadable at https://www.geoda.uiuc.edu/) and the robust LM tests are also conducted by GEODA. As we can observe in Figure 5.3, the spatial extent of houses in one cluster varies from cluster to cluster. Since the differences in the distribution patterns differ significantly, we expect the cutoff distance of the best fit weight matrices may differ from cluster to cluster.

The robust LM test results are listed in Table 6.3. The preferred model is chosen as follows. If one is significant and the other is not, then the significant model is selected. If both models are significant, the one with the higher test statistics is chosen. For example, for the 1+2 cluster with the 800 meter weight matrix, the statistics for the lag model is not significant (0.76) while it is statistically significant at less than the 1 % level for the error model. In such cases, we choose the spatial error model simply based on the test outcome. For the 1+2 cluster with the 200 meter weight matrix, the test shows that both the lag and the error models are statistically significant at the 9 % level for the lag model and less than the 1 % level for the error model and the value of the statistics are 3.05 and 46.26, respectively. Therefore, for the case both the lag and the error models are statistically significant, we choose the model with higher value, which is in this case the spatial error model. For all clusters and all weight matrices case, the spatial error specification is preferred. Therefore, we proceed with the spatial error specifications for all clusters.

The spatial error models with all four weight matrices are estimated with GMM. GMM is implemented with MATLAB by using the Kelejian and Prucha's way discussed in section 2.3.3. We compared the estimated adjusted R-square between the models with

different weight matrices and choose the model with the weight matrix with the highest adjusted R-square. For the individual houses case, most of the preferred models are with the weight matrix with the 400 meter cutoff distance. Clusters that do not have the 400 meter weight matrix are cluster 6 (800 meter), cluster 8 (1600 meter) and cluster 10 (800 meter). Therefore, the following analysis includes spatial error models with the preferred weight matrices for each cluster.

		Cluster									
	W	1+2	3	4	5	6	7	8	9	10	11
Lag	200	3.05	1.57	5.28	14.19	2.22	9.83	16.72	0.85	6.08	0.96
		(0.08)	(0.21)	(0.02)	(0.00)	(0.14)	(0.00)	(0.00)	(0.36)	(0.01)	(0.33)
Error		46.26	18.72	37.40	122.14	188.90	53.29	36.97	45.00	134.32	15.17
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lag	400	5.49	2.83	1.62	7.97	1.32	3.97	8.71	1.39	1.36	7.51
		(0.02)	(0.09)	(0.20)	(0.00)	(0.25)	(0.05)	(0.00)	(0.24)	(0.24)	(0.01)
Error		68.13	32.10	54.28	171.90	383.87	44.95	63.13	34.01	221.05	59.60
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lag	800	0.09	0.53	0.95	77.35	1.70	3.61	8.71	0.01	7.49	0.78
		(0.76)	(0.47)	(0.33)	(0.00)	(0.19)	(0.06)	(0.00)	(0.91)	(0.01)	(0.38)
Error		31.28	28.83	53.42	168.49	404.40	33.94	63.13	13.46	413.96	41.32
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lag	1600	0.36	4.67	2.25	57.97	4.50	2.41	1.74	0.12	1.00	0.28
		(0.55)	(0.03)	(0.13)	(0.00)	(0.03)	(0.12)	(0.19)	(0.73)	(0.32)	(0.59)
Error		50.43	7.02	101.12	130.22	222.17	38.29	66.69	6.56	113.23	5.04
		(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.02)

* ( ) probability

Table 6.3. Robust LM Test Result: Individual Houses Case

Estimated results of GMM are reported in Table 6.4. In the third row, the cutoff distance of the preferred weight matrix is listed. The magnitudes of the coefficients estimated are not too different from the ones from the OLS estimation. However, the degree of statistical significance increased after controlling for the spatial effects. For example, the age squared term for cluster 6 is significant at the 10 percent level and the deck variable is significant at least at the 10 percent level for all clusters while they are not in the OLS estimation. As for the fecal coliform variable, two clusters, 7 and 10 which have positive significant outcome in OLS are not significant at the 10 percent level in the GMM estimation. The estimated spatial error coefficients are all positive significant at the one percent level.

The fecal coliform variable is estimated as positive and significant, in other words, the higher the fecal coliform counts in the Lake, the higher the housing value is, for Cluster 4 and 8. Cluster 4 is located in Lorain County, not along the coast line, however relatively close to the Lake. The average value of the fecal variable is 292 counts per ml. Considering that the overall mean is 255, we cannot observe any extreme condition in terms of fecal in this cluster. It is possible that we are omitting a variable which is correlated with the level of fecal and is specific to this cluster. Interaction terms between the distance to the closest beach and each water quality variable are dropped from the model because of the mixed signs observed for the coefficients estimated for the water quality variables.

The mean fecal value for Cluster 8 is the lowest among others, 87 counts per 100 ml. If there is a certain factor which costs home owners in the effort of reducing fecal coliform or organic matter discharges, having very low fecal may be costing them higher than they are willing. If this is the reason, it is possible to have positive fecal coliform coefficient. It is also possible that we omitted a variable which is related with the fecal coliform counts in the more general sense.

	Cluster									
	3		4		5		6		7	
W	400		400		800		400		400	
CONSTANT	7.734	***	7.226	***	7.802	***	7.228	***	9.083	***
	(17.69)		(21.19)		(34.47)		(30.69)		(26.50)	
LNLOTACR	0.042	**	0.065	***	0.150	***	0.092	***	0.026	**
	(2.09)		(5.43)		(10.20)		(6.93)		(2.40)	
LNBLDGSF	0.414	***	0.413	***	0.368	***	0.409	***	0.319	***
	(9.55)		(11.30)		(14.56)		(14.87)		(9.11)	
BATHN	0.013		0.080	***	0.092	***	0.091	***	0.086	***
	(0.58)		(4.33)		(6.91)		(5.01)		(3.54)	
GRGSQF	0.000		0.000		0.000		0.000	***	0.000	***
	(0.42)		(-0.02)		(1.37)		(4.58)		(5.69)	
AGE	-0.009	***	-0.007	***	-0.008	***	-0.005	***	-0.006	***
	(-7.28)		(-7.74)		(-6.53)		(-5.21)		(-4.82)	
AGE2	0.000	***	0.000	***	0.000	***	0.000	**	0.000	***
	(4.84)		(5.68)		(4.78)		(2.14)		(4.38)	
AIRCNDD	0.194	*	0.190	***	0.160	***	0.055	***	0.125	***
	(1.90)		(4.18)		(3.87)		(3.15)		(4.69)	
DECKD	0.079	***	0.047	*	0.035		0.071	***	0.048	*
	(2.85)		(1.65)		(1.64)		(3.45)		(1.72)	
FIREPLD	0.120	***	0.081	***	0.077	***	0.061	***	0.135	***
	(5.01)		(4.51)		(5.70)		(3.42)		(6.02)	
SDRANK	-0.005		-0.003	***	-0.003	***	-0.001		0.003	*
	(-1.47)		(-3.03)		(-3.23)		(-0.87)		(1.82)	
LNBEACH	0.160	*	0.052		-0.031	**	-0.056	***	-0.126	***
	(1.87)		(0.95)		(-2.52)		(-4.01)		(-10.22)	
LNFECAL	0.000		0.023	**	-0.008		-0.014	*	0.006	
	(-0.00)		(2.01)		(-1.00)		(-1.72)		(0.84)	
LNSECCHI	0.039		0.101	***	0.012		0.163	***	-0.007	
	(1.06)		(3.85)		(0.55)		(6.16)		(-0.16)	
lambda	0.258	***	0.272	***	0.393	***	0.392	***	0.291	***
	(4.33)		(6.86)		(11.15)		(15.60)		(6.47)	
AdjR2	0.675		0.691		0.687		0.667		0.532	
N	544		839		1229		1152		693	

Table 6.4. GMM Result for Each Cluster: Individual Houses Case

Table 6.4 (continued)

	Cluster									
	8		9		10		11		1+2	
W	1600		400		800		400		400	
CONSTANT	5.912	***	8.529	***	7.410	***	6.838	***	7.875	***
	(21.16)		(34.18)		(25.78)		(25.49)		(33.53)	
LNLOTACR	0.057	***	0.068	***	0.095	***	0.130	***	0.079	***
	(5.72)		(8.69)		(7.24)		(9.78)		(13.26)	
LNBLDGSF	0.489	***	0.337	***	0.411	***	0.625	***	0.424	***
	(19.12)		(12.68)		(18.66)		(20.94)		(17.71)	
BATHN	0.055	***	0.078	***	0.065	***	0.082	***	0.050	***
	(3.02)		(6.91)		(5.60)		(5.30)		(3.95)	
GRGSQF	0.000	***	0.000		0.000		0.000		0.000	
	(4.00)		(0.57)		(-0.98)		(-0.50)		(-1.15)	
AGE	-0.006	***	-0.010	***	-0.001		-0.008	***	-0.007	***
	(-5.60)		(-13.39)		(-1.15)		(-7.24)		(-11.25)	
AGE2	0.000	***	0.000	***	0.000		0.000	***	0.000	***
	(2.66)		(10.73)		(-0.33)		(3.79)		(7.47)	
AIRCNDD	0.027		0.105	***	0.086	***	0.130	***	0.126	***
	(1.53)		(4.47)		(3.26)		(4.57)		(5.22)	
DECKD	0.034	*	0.048	**	0.068	***	0.085	***	0.110	***
	(1.67)		(2.55)		(3.61)		(4.66)		(5.51)	
FIREPLD	0.105	***	0.078	***	0.070	***	0.069	***	0.087	***
	(5.81)		(6.85)		(5.57)		(3.89)		(6.82)	
SDRANK	-0.005	**	0.014	***	0.009	***	-0.011	***	0.000	
	(-2.02)		(6.11)		(2.65)		(-6.33)		(-0.67)	
LNBEACH	0.167	***	-0.083	*	-0.112	**	0.024	**	-0.104	***
	(3.60)		(-1.70)		(-1.96)		(2.13)		(-3.68)	
LNFECAL	0.063	***	-0.017	**	0.008		-0.067	***	0.001	
	(7.99)		(-2.34)		(0.95)		(-5.39)		(0.12)	
LNSECCHI	0.183	***	0.019		0.052	**	-0.030		0.050	**
	(6.36)		(1.25)		(2.42)		(-1.46)		(2.29)	
lambda	0.345	***	0.207	***	0.410	***	0.175	***	0.214	***
	(8.73)		(5.50)		(10.85)		(4.27)		(7.73)	
AdjR2	0.669		0.671		0.685		0.849		0.618	
N	971		1185		1334		1218		1500	

6.2.3 Estimated Marginal Implicit Prices

Based on the estimated coefficient of the GMM models, marginal implicit prices (MIP) are computed for individual observations by using individual variables and estimated coefficients. For each cluster, we used the coefficients estimated for the cluster and all MIP computed are merged in order to proceed to the second stage of the hedonic estimation. MIPs are computed as $\hat{\beta} * P^i$ for the variables that are entered without taking natural logarithm. It is $\hat{\beta} * \frac{P^i}{Q^i}$ for the variables with logarithm where $\hat{\beta}$ is the estimated coefficient of the GMM model, P^i is the house price for the *i* th observation, and Q^i is the quantity of the relevant variable for the *i* th observation. The individual values are used to compute the individual specific MIPs.

Table 6.5 lists the descriptive statistics of the computed MIP for all data combined. All prices are expressed in 1996 dollars. Marginal implicit prices are interpreted as one unit increase in a certain variable from the current state will increase the house owner's willingness to pay by the MIP amount. For example, one year increase in a certain house's age will decrease the housing price by 552 dollars. If school district ranking increases by one, housing price will increase by 89 dollars. This is much smaller than we expected. If the house locates one kilometer away from the current place, the house value will decrease by 707 dollars. We expect the fecal coliform variable to be negative. However, since we have mixed signs in the GMM estimation, this turns out to be positive. As for the secchi disk depth readings, one centimeter increase in the lake water clarity increases the housing value by 30 dollars.

As we mentioned earlier, some of the fecal coliform counts variable have unexpected signs. However, we believe that the bacterial counts should have negative influence on housing values although it may not be significantly influencing. In order to further analyze the influence of the fecal coliform counts, we select the clusters with negative coefficients of the fecal coliform counts and implemented a separate analysis. The result of the analysis will reflect only the houses whose price is affected negatively by the fecal coliform.

Furthermore, we estimate the observations with housing values which are both significantly and negatively influenced by the fecal coliform counts. The result of the analysis with this set of data will represent the houses which are influenced by the fecal coliform negatively and significantly. We call the first type of the observations (with negative, but not necessary significant coefficients on the fecal coliform counts) "Correct signed" or COR, and the second type (with significantly negative coefficients) as "Significant" or SIG in the following analysis. COR data include clusters 3, 5, 6, 9 and 11, total of 5,238 observations, and SIG data contains clusters 6, 9 and 11, including 3,555 houses. MIPs computed from COR and SIG data are shown in Table 6.6.

	MIP for All Data (in 1996$)			
	Mean	St.Dev.	Min.	Max.
PLOTACR	37.39	31.69	0.03	636.32
PBLDGSF	29.14	12.39	5.31	192.02
PBATHN	7941.53	5146.55	660.23	55098.79
PGRGSQF	4.57	9.87	-19.10	89.63
PAGE	-552.60	535.40	-5061.19	2498.99
PAIRCNDD	13281.61	9736.34	1354.25	86909.57
PDECKD	7539.45	5682.78	1703.80	57169.58
PFIREPLD	9227.69	5005.22	3069.45	53315.20
PSDRANK	-89.42	974.28	-7095.83	4675.38
PBEACH	-707.08	15720.96	-1378243.50	61760.77
PFECAL	7.52	79.92	-480.44	1279.02
PSECCHI	30.34	47.69	-96.65	649.47
ADJINC	9775.55	22732.22	-173529.05	415416.15
N	10665			

Table 6.5. Estimated Marginal Implicit Prices for All Data: Individual Houses Case

MIPs estimated by including only COR reveals that an increase in one fecal coliform decreases housing value by 21.6 dollars. This is MIP for the houses which have negative influence from the bacterial counts. MIP for the fecal coliform for the houses negatively significantly influenced by the fecal coliform is - 30.5, meaning the increase in one fecal coliform count will decrease the housing value by 30.5 dollars. It is expected to have larger value for the SIG data than the COR data since the houses that are significantly affected by the fecal coliform should have higher MIP values compared to the data which include both significant and insignificant values.

	MIP for Fecal COR Data (in 1996$)				MIP for Fecal SIG Data (in 1996$)			
	Mean	St.Dev.	Min.	Max.	Mean	St.Dev.	Min.	Max.
PLOTACR	49.3	36.2	0.1	636.3	49.5	39.0	0.1	636.3
PBLDGSF	31.0	14.8	5.3	187.9	34.2	16.5	8.0	187.9
PBATHN	9345.9	6143.7	660.2	55098.8	11005.3	6358.4	3883.4	55098.8
PGRGSQF	7.0	9.2	-15.4	89.6	5.1	10.2	-15.4	89.6
PAGE	-697.9	606.3	-5061.2	2499.0	-797.8	623.3	-5061.2	2499.0
PAIRCNDD	14882.7	10290.8	2771.5	86909.6	13607.6	10496.1	2771.5	86909.6
PDECKD	8040.3	6464.8	1756.7	57169.6	9410.4	6913.4	2390.0	57169.6
PFIREPLD	9221.7	5468.1	3069.5	45871.9	9112.0	5146.0	3069.5	45871.9
PSDRANK	-228.0	1268.5	-7095.8	4675.4	-165.8	1539.2	-7095.8	4675.4
PBEACH	-707.5	19709.6	-1378243	61760.8	-971.2	23959.8	-1378243	61760.8
PFECAL	-21.6	39.5	-480.4	0.0	-30.5	45.7	-480.4	-0.3
PSECCHI	22.0	52.9	-96.6	516.7	27.7	63.5	-96.6	516.7
ADJINC	11172	22757	-173529	415416	15723	24657	-173529	415416
N	5238				3555			

Table 6.6. Estimated Marginal Implicit Prices for Fecal COR and Fecal SIG Data: Individual Houses Case

	MIP for Secchi COR Data (in 1996$)				MIP for Secchi SIG Data (in 1996$)			
	Mean	St.Dev.	Min.	Max.	Mean	St.Dev.	Min.	Max.
PLOTACR	34.7	27.9	0.1	632.9	31.5	26.8	0.1	632.9
PBLDGSF	26.7	8.9	5.3	192.0	28.3	9.5	5.9	192.0
PBATHN	6996.0	3962.6	660.2	53226.2	7001.5	4124.1	2521.3	53226.2
PGRGSQF	4.5	9.0	-19.1	89.6	2.8	10.0	-19.1	89.6
PAGE	-500.4	433.7	-3393.6	2499.0	-441.2	385.5	-3063.6	1565.4
PAIRCNDD	12039.4	8794.6	1354.3	73602.5	10520.8	8751.7	1354.3	67575.5
PDECKD	6718.4	4527.5	1703.8	51384.2	7466.9	4818.0	1703.8	51384.2
PFIREPLD	8534.2	4403.0	3069.5	53315.2	8280.0	4071.2	3069.5	53315.2
PSDRANK	129.5	714.7	-2778.4	4675.4	-8.3	475.0	-2778.4	2850.3
PBEACH	-732	15298	-1378243	7671	-919	18667	-1378243	7671
PFECAL	15.1	81.4	-272.9	1279.0	26.6	97.9	-272.9	1279.0
PSECCHI	40.5	46.3	1.5	649.5	56.0	49.9	6.0	649.5
ADJINC	8111.5	19887	-173529.1	415416.2	9595.6	21280.5	-173529.1	228782.6
N	8754				5796			

Table 6.7. Estimated Marginal Implicit Prices for Secchi COR and Secchi SIG Data: Individual Houses Case

We also construct the data for the secchi variable in the same way as we do with the fecal. The COR data for secchi include the observations in the cluster with positive coefficients for the secchi variable. All clusters except for cluster 7 and 11 are included in this set of data and the total number of observations is 8,754. The SIG data is constructed with five clusters, 4, 6, 8, 10 and 1+2. The estimated MIPs are listed in Table 6.7. For houses in the COR data, one centimeter increase in water clarity increases the housing value by 40.5 dollars while it is 56 dollars for the observations in the SIG data.

6.3 Estimated Results: Census Block Group Case

In this section, the result of the first stage estimation using the clustering with census block group is reported. Given the results of clustering with census block groups, we assigned individual houses data to each cluster. Therefore, data included in the analysis is the same as the individual houses cases.

6.3.1 Estimated Result of OLS

Lot acreage is positive significant for all but cluster 7 while building square feet is significant for all clusters. Number of bathrooms is also significant for all but cluster 7. Interestingly, the estimated garage square feet coefficients have mixed result. Five clusters have positive significant outcome while three clusters have negative significant result. The clusters with negative coefficients for garage square feet have significantly smaller garage size comparing to the ones with positive significant coefficients. Age of the house variable is significant at least at the five percent level while age squares is

significant all but for cluster 7. Clusters 1, 3, 6 and 9 have expected signs on school district ranking while clusters 2, 4 and 10 have unexpected positive signs. The air-conditioning variable is positive significant for all but cluster 4 while the deck dummy is significant except for clusters 4 and 9. The fireplace dummy variable is statistically significant for all clusters.

Distance to the closest beach is negatively significant for clusters 1, 2, 7, 8 and 9, indicating that housing price of the relevant houses have an inverse relation with the distance from the beach. On the other hand, cluster 6 has positive significant coefficient for the value, indicating the higher the housing price is, the more the distance from the closest beach is. The fecal coliform variable is negative significant in two clusters, 2 and 3, while it is positive significant in clusters 5, 6 and 10. Secchi disk reading is positive significant for clusters 1, 2, 6, 8 and 10. The signs are consistently positive for secchi values across all clusters.

	Cluster 1		Cluster 2		Cluster 3		Cluster 4		Cluster 5	
CONSTANT	7.820	***	7.077	***	6.434	***	7.434	***	8.670	***
	(47.46)		(23.03)		(27.76)		(9.78)		(16.20)	
LNLOTACR	0.077	***	0.052	***	0.111	***	-0.080	**	0.055	***
	(12.86)		(4.24)		(8.39)		(-2.52)		(4.06)	
LNBLDGSF	0.423	***	0.442	***	0.628	***	0.451	***	0.311	***
	(21.66)		(12.09)		(21.85)		(5.46)		(7.17)	
BATHN	0.068	***	0.079	***	0.138	***	0.151	***	0.087	***
	(7.12)		(3.26)		(8.64)		(3.19)		(2.80)	
GRGSQF	-0.0002	***	0.0002	***	0.000004		0.0002	*	0.0002	***
	(-5.02)		(6.04)		(0.11)		(1.86)		(4.85)	
AGE	-0.008	***	-0.005	***	-0.006	***	-0.006	**	-0.005	***
	(-16.42)		(-4.10)		(-7.45)		(-2.01)		(-3.27)	
AGE2	0.0001	***	0.00002	**	0.00003	***	0.0001	***	0.00003	**
	(12.13)		(2.56)		(3.89)		(2.65)		(2.45)	
AIRCNDD	0.159	***	0.080	***	0.074	***	0.068		0.130	***
	(7.17)		(3.52)		(3.37)		(1.26)		(4.25)	
DECKD	0.069	***	0.080	***	0.086	***	0.016		0.091	**
	(4.76)		(2.97)		(4.56)		(0.30)		(2.30)	
FIREPLD	0.103	***	0.086	***	0.084	***	0.209	***	0.101	***
	(10.66)		(3.76)		(4.85)		(4.56)		(3.73)	
SDRANK	-0.006	***	0.005	***	-0.007	***	0.017	***	-0.002	
	(-13.73)		(4.06)		(-5.04)		(3.24)		(-1.25)	
LNBEACH	-0.042	***	-0.074	***	-0.011		-0.036		-0.049	
	(-2.82)		(-5.29)		(-1.19)		(-1.47)		(-0.65)	
LNFECAL	-0.001		-0.050	***	-0.048	***	-0.003		0.016	**
	(-0.28)		(-5.46)		(-5.87)		(-0.16)		(2.07)	
LNSECCHI	0.040	***	0.205	***	0.030		0.124		0.015	
	(2.94)		(5.93)		(1.45)		(1.27)		(0.29)	
AdjR2	0.70		0.54		0.80		0.41		0.55	
N	2628		781		1494		280		323	

Table 6.8. Estimated Result of OLS for Each Cluster: CBG Case

Table 6.8 (continued)

	Cluster 6		Cluster 7		Cluster 8		Cluster 9		Cluster 10	
CONSTANT	5.802	***	8.421	***	7.788	***	7.322	***	6.736	***
	(22.78)		(15.36)		(30.33)		(29.81)		(27.03)	
LNLOTACR	0.058	***	0.030		0.078	***	0.164	***	0.117	***
	(6.90)		(1.17)		(13.55)		(11.16)		(9.72)	
LNBLDGSF	0.504	***	0.412	***	0.436	***	0.415	***	0.446	***
	(21.58)		(8.42)		(17.08)		(15.04)		(20.16)	
BATHN	0.071	***	0.042		0.044	***	0.101	***	0.068	***
	(4.22)		(0.89)		(3.19)		(6.94)		(5.86)	
GRGSQF	0.0001	***	0.0002	**	-0.000069	*	0.0002		-0.0001	*
	(4.19)		(2.21)		(-1.71)		(1.62)		(-1.69)	
AGE	-0.005	***	-0.004	**	-0.008	***	-0.010	***	-0.007	***
	(-5.02)		(-2.17)		(-9.83)		(-8.01)		(-9.14)	
AGE2	0.0000	*	0.00002		0.00005	***	0.0001	***	0.00005	***
	(1.77)		(1.03)		(5.95)		(6.00)		(5.24)	
AIRCNDD	0.067	***	0.112	***	0.126	***	0.206	***	0.121	***
	(4.08)		(3.21)		(4.54)		(4.23)		(4.34)	
DECKD	0.043	**	0.108	*	0.104	***	0.025		0.075	***
	(2.17)		(1.71)		(4.48)		(1.11)		(3.92)	
FIREPLD	0.110	***	0.098	**	0.089	***	0.088	***	0.092	***
	(6.70)		(2.42)		(6.33)		(6.07)		(7.45)	
SDRANK	-0.006	***	0.003		0.000		-0.002	***	0.003	**
	(-3.71)		(0.55)		(0.58)		(-3.52)		(1.84)	
LNBEACH	0.225	***	-0.460	***	-0.111	***	-0.025	**	0.016	
	(4.86)		(-4.25)		(-4.01)		(-2.58)		(0.51)	
LNFECAL	0.053	***	0.066		-0.004		-0.004		0.021	***
	(7.57)		(1.19)		(-0.47)		(-0.43)		(2.58)	
LNSECCHI	0.149	***	0.037		0.060	**	0.012		0.079	***
	(5.89)		(0.63)		(2.53)		(0.48)		(3.50)	
AdjR2	0.66		0.44		0.60		0.65		0.70	
N	1118		251		1284		1124		1382	

As in the case with individual houses, we test whether the estimated coefficients are significantly different from each other for every cluster. The chow test result is listed in Table 6.9. There is no cluster with an F value less than 2. Therefore, we do not merge any clusters and treat each cluster as a separate submarket.

	Cluster									
	1	2	3	4	5	6	7	8	9	10
1	-	20.50	34.07	14.04	8.60	17.56	8.97	5.84	7.20	5.59
2		-	18.11	6.62	3.32	9.55	4.52	5.94	15.51	12.14
3			-	14.54	8.93	14.18	4.84	24.73	16.94	17.43
4				-	5.59	13.86	4.41	8.26	14.22	10.65
5					-	5.15	1.86	4.39	5.05	4.97
6						-	4.66	9.00	9.06	9.42
7							-	2.54	6.37	4.90
8								-	6.49	2.44
9									-	4.83
10										-

Table 6.9. Chow Test Result: CBG Case

6.3.2 Estimated Result of Spatial Hedonic Model

In order to estimate the spatial hedonic models, we first implemented robust LM test by using four different weight specifications as in the case of individual houses. Except for cluster 5 with weight matrix of 1600 meter cutoff distance (Lag model is preferred) and cluster 7 with weight matrix of 1600 meter cutoff distance (spatial models are not significant), in all clusters and weight matrices the spatial error model is the preferred model. We compare the adjusted R-squares with the preferred models among different

weight specifications and choose the highest R-squares valued model for the use of the following analysis. As for clusters 5 and 7, spatial error specification with weight matrices of 800 meter and 400 meter are chosen after the comparison. Therefore, the spatial error model is used for all clusters. Estimated GMM results can be found in Table 6.11 with the selected weight cutoff values expressed in the third row.

		Cluster									
	W	1	2	3	4	5	6	7	8	9	10
Lag	200	5.00	5.92	3.11	1.41	0.24	18.32	0.31	1.93	17.79	13.94
		(0.03)	(0.01)	(0.08)	(0.23)	(0.62)	(0.00)	(0.58)	(0.16)	(0.00)	(0.00)
Error		195.75	63.53	261.87	47.08	6.38	36.67	24.91	17.88	99.95	199.40
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lag	400	4.44	17.15	3.77	0.24	2.57	15.94	0.28	3.65	6.34	3.95
		(0.04)	(0.00)	(0.05)	(0.63)	(0.11)	(0.00)	(0.60)	(0.06)	(0.01)	(0.05)
Error		261.26	53.64	793.97	45.42	7.27	80.89	31.35	25.01	140.05	277.88
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lag	800	0.99	10.31	5.19	0.30	2.01	12.28	3.42	0.10	71.58	2.86
		(0.32)	(0.00)	(0.02)	(0.58)	(0.16)	(0.00)	(0.06)	(0.75)	(0.00)	(0.09)
Error		269.59	87.61	736.58	39.47	7.87	70.25	37.86	10.35	131.92	424.87
		(0.00)	(0.00)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Lag	1600	4.08	7.61	4.71	0.15	11.06	6.52	0.01	0.01	58.53	1.79
		(0.04)	(0.01)	(0.03)	(0.70)	(0.00)	(0.01)	(0.90)	(0.94)	(0.00)	(0.18)
Error		499.62	104.15	282.26	40.72	0.24	61.40	0.35	34.82	99.33	27.22
		(0.00)	(0.00)	(0.00)	(0.00)	(0.63)	(0.00)	(0.55)	(0.00)	(0.00)	(0.00)

Table 6.10. Robust LM Test Result: CBG Case

The magnitudes and signs of estimated coefficients do not change significantly between OLS and GMM. Although for the case of fecal coliform in cluster 10 is positive significance for the individual case, it is not significant for the GMM case. The spatial error coefficients are all positive significant at the 1 percent level.

We have unexpected positive significant signs for the fecal variable for Cluster 5 and Cluster 6. Cluster 6 is almost identical to Cluster 8 in the individual houses case which also has a positive significant result for fecal. Cluster 6 has a very high mean fecal value among others, 515 counts per 100 ml while the overall average is 255. The area this cluster covers has a very high percentage of agricultural land. Since one source of fecal coliform is organic discharges from farm lands, it may be possible to say that many house owners of this area actually benefit from discharging organic matters into streams. We cannot prove how this possibility may affect their housing price with data at hand, but again, it is possible to assume that there is omitted variable which is related to fecal and is causing the positive effect on the housing price.

	Cluster									
	1		2		3		4		5	
W	400		1600		400		400		800	
CONSTANT	8.119	***	7.400	***	6.980	***	7.679	***	8.582	***
	(48.29)		(25.21)		(30.29)		(10.91)		(15.83)	
LNLOTACR	0.072	***	0.060	***	0.122	***	-0.068	**	0.061	***
	(10.86)		(4.74)		(9.07)		(-2.16)		(4.30)	
LNBLDGSF	0.389	***	0.402	***	0.547	***	0.482	***	0.318	***
	(19.97)		(11.83)		(19.63)		(6.26)		(7.74)	
BATHN	0.062	***	0.061	***	0.115	***	0.085	**	0.076	***
	(6.65)		(2.71)		(7.52)		(1.99)		(2.58)	
GRGSQF	0.000	***	0.000	***	0.000		0.000	*	0.000	***
	(-5.55)		(5.35)		(-0.30)		(1.69)		(5.07)	
AGE	-0.008	***	-0.007	***	-0.006	***	-0.006	**	-0.004	***
	(-15.00)		(-5.61)		(-6.50)		(-2.30)		(-2.70)	
AGE2	0.000	***	0.000	***	0.000	**	0.000	***	0.000	**
	(11.07)		(3.77)		(2.36)		(2.59)		(2.08)	
AIRCNDD	0.154	***	0.071	***	0.080	***	0.038		0.129	***
	(7.31)		(3.17)		(3.69)		(0.79)		(4.40)	
DECKD	0.073	***	0.067	***	0.091	***	0.023		0.090	**
	(5.27)		(2.68)		(5.25)		(0.49)		(2.40)	
FIREPLD	0.094	***	0.064	***	0.080	***	0.132	***	0.109	***
	(10.05)		(2.97)		(4.81)		(3.07)		(4.18)	
SDRANK	-0.006	***	0.003	**	-0.008	***	0.017	**	-0.002	
	(-10.48)		(2.21)		(-4.63)		(2.56)		(-0.82)	
LNBEACH	-0.047	**	-0.082	***	0.002		-0.061	**	-0.042	
	(-2.39)		(-4.57)		(0.19)		(-2.07)		(-0.48)	
LNFECAL	0.000		-0.039	***	-0.038	***	0.001		0.013	*
	(-0.08)		(-4.28)		(-4.30)		(0.04)		(1.65)	
LNSECCHI	0.041	***	0.204	***	0.026		0.056		0.007	
	(3.10)		(6.06)		(1.33)		(0.59)		(0.15)	
lambda	0.302	***	0.346	***	0.372	***	0.358	***	0.230	***
	(13.13)		(9.14)		(13.62)		(6.92)		(3.32)	
AdjR2	0.73		0.59		0.82		0.50		0.57	
N	2628		781		1494		280		323	

Table 6.11. GMM Result for Each Cluster: CBG Case

Table 6.11 (continued)

	Cluster									
	6		7		8		9		10	
W	400		400		1600		800		800	
CONSTANT	6.150	***	8.569	***	7.885	***	7.610	***	7.392	***
	(22.63)		(15.67)		(30.27)		(31.74)		(27.76)	
LNLOTACR	0.058	***	0.032		0.083	***	0.152	***	0.101	***
	(6.64)		(1.27)		(13.07)		(9.70)		(7.83)	
LNBLDGSF	0.470	***	0.390	***	0.423	***	0.385	***	0.407	***
	(19.89)		(8.57)		(16.68)		(14.32)		(18.45)	
BATHN	0.065	***	0.008		0.039	***	0.093	***	0.070	***
	(3.98)		(0.19)		(2.85)		(6.74)		(6.27)	
GRGSQF	0.000	***	0.000	**	0.000	*	0.000		0.000	**
	(3.72)		(2.37)		(-1.94)		(1.48)		(-2.20)	
AGE	-0.005	***	-0.005	**	-0.008	***	-0.008	***	-0.007	***
	(-5.42)		(-2.54)		(-9.45)		(-5.95)		(-7.46)	
AGE2	0.000	**	0.000		0.000	***	0.000	***	0.000	***
	(2.44)		(1.62)		(5.66)		(4.36)		(4.43)	
AIRCNDD	0.057	***	0.090	***	0.128	***	0.177	***	0.102	***
	(3.54)		(2.80)		(4.68)		(3.81)		(3.89)	
DECKD	0.039	**	0.059		0.104	***	0.042	*	0.066	***
	(2.04)		(1.00)		(4.55)		(1.90)		(3.74)	
FIREPLD	0.098	***	0.089	**	0.084	***	0.079	***	0.071	***
	(6.09)		(2.38)		(5.99)		(5.68)		(5.86)	
SDRANK	-0.004	**	0.002		0.001		-0.003	***	0.003	*
	(-2.26)		(0.44)		(0.56)		(-3.04)		(1.82)	
LNBEACH	0.188	***	-0.460	***	-0.114	***	-0.024	*	0.008	
	(3.35)		(-3.23)		(-3.44)		(-1.80)		(0.16)	
LNFECAL	0.051	***	0.065		-0.003		-0.006		0.008	
	(7.31)		(1.05)		(-0.30)		(-0.76)		(0.98)	
LNSECCHI	0.148	***	0.053		0.055	**	0.016		0.047	**
	(5.94)		(0.99)		(2.39)		(0.68)		(2.22)	
lambda	0.250	***	0.352	***	0.199	***	0.370	***	0.450	***
	(7.34)		(5.25)		(5.60)		(10.04)		(12.39)	
AdjR2	0.67		0.50		0.61		0.68		0.73	
N	1118		251		1284		1124		1382	

6.3.3 Estimated Marginal Implicit Prices

Computed MIPs are listed in Table 6.13. Compared to the results from the individual houses case, we observe that the differences of marginal price for lot acreage, building square feet, bathroom, age, air-conditioning and fireplace are less than five percent. On the other hand, school district ranking has the largest difference and mark up to 245 percent in magnitude. For the census block group case, if the school district ranking increases by one rank, the housing price is suggested to increase by 308 dollars. Marginal price for the distance to the beach is 10 percent lower for the census block group case where one kilo meter increase in the distance from the closest beach will decrease the housing value by 632 dollars. MIP for fecal coliform is again positive due to the positive signed estimates in the GMM estimation. As for the secchi depth readings, an increase in the water clarity by one centimeter will increase the housing value by 34 dollars, which is about 12 percent higher than the case for individual houses.

MIPs by including the expected signed observations (COR data) are also computed and listed in Table 6.14 together with MIPs derived by including observations which have statistically significant outcome with expected signs (SIG data). MIP of fecal is derived as 18 dollars for one count decrease in fecal coliform counts per 100 ml for the COR data while it is 58 dollars for the SIG data. If we compare with the individual houses case (21.6 dollars for the COR and 30.5 dollars for the SIG data), it is lower for the COR data and is higher for the SIG data.

The same set of results are provided for the secchi variable in Table 6.15. One centimeter increase in the water clarity will increase the value of the house by 34 dollars if the houses are affected positively by the secchi readings regardless of the significance. It is 43 dollars for the houses significantly influenced by the lake water clarity. The amount estimated is lower than the individual houses case (40.5 dollars for the COR data and 56 dollars for the SIG data).

	MIP for All Data (in 1996$)			
	Mean	St.Dev.	Min.	Max.
PLOTACR	35.47	36.14	-413.07	838.63
PBLDGSF	29.17	11.52	4.98	251.10
PBATHN	8223.38	6635.20	406.30	76797.24
PGRGSQF	-3.58	16.98	-79.78	115.82
PAGE	-529.54	447.04	-3728.63	2839.52
PAIRCNDD	12638.89	7373.93	1878.70	59671.01
PDECKD	8142.45	5790.71	1134.65	61195.37
PFIREPLD	9510.43	5071.53	3199.95	53267.61
PSDRANK	-308.99	729.54	-5474.81	6282.84
PBEACH	-631.85	4077.61	-251114	57555.15
PFECAL	3.76	76.57	-944.25	1036.81
PSECCHI	33.93	41.88	1.59	645.14
ADJINC	15625.97	25216.22	-241102	280592.15
	10665			

Table 6.12 Estimated Marginal Implicit Prices for All Data: CBG Case

	MIP for Fecal COR Data (in 1996$)				MIP for Fecal SIG Data (in 1996$)			
	Mean	St.Dev.	Min.	Max.	Mean	St.Dev.	Min.	Max.
PLOTACR	40.1	35.2	0.1	838.6	53.1	47.9	0.1	838.6
PBLDGSF	30.5	11.9	5.0	251.1	39.2	15.0	7.1	251.1
PBATHN	9279.9	7505.2	1970.7	76797.2	14846.8	10706.3	3042.9	76797.2
PGRGSQF	-6.4	18.1	-79.8	115.8	6.4	13.4	-8.0	115.8
PAGE	-636.5	468.5	-3728.6	1512.3	-720.7	520.2	-3728.6	1023.5
PAIRCNDD	15026.7	7352.3	3534.4	59671.0	11444.2	7031.1	3534.4	53550.7
PDECKD	9618.8	6234.4	2082.6	61195.4	12546.0	8188.4	3350.6	61195.4
PFIREPLD	10183.3	5339.5	3200.0	53267.6	11139.7	7054.6	3200.0	53267.6
PSDRANK	-515.5	681.2	-5474.8	1858.3	-767.4	1032.9	-5474.8	1858.3
PBEACH	-597.3	1856.1	-75984	57555.2	-461.8	2171.7	-23057	57555.2
PFECAL	-18.0	52.0	-944.3	0.0	-53.6	82.7	-944.3	-0.7
PSECCHI	32.0	44.5	2.2	645.1	58.3	70.7	3.6	645.1
ADJINC	14044.0	25468.6	-120485	202278.4	34347.0	29820.6	-120485	202278.4
	7311				2275			

Table 6.13 Estimated Marginal Implicit Prices for Fecal COR and Fecal SIG Data: CBG Case

	MIP for Secchi COR Data (in 1996$)				MIP for Secchi SIG Data (in 1996$)			
	Mean	St.Dev.	Min.	Max.	Mean	St.Dev.	Min.	Max.
PLOTACR	35.5	36.1	-413.1	838.6	28.6	21.1	0.1	562.7
PBLDGSF	29.2	11.5	5.0	251.1	27.4	8.2	5.0	184.3
PBATHN	8223.4	6635.2	406.3	76797.2	6307.8	2839.5	1970.7	35419.4
PGRGSQF	-3.6	17.0	-79.8	115.8	-8.7	17.5	-79.8	115.8
PAGE	-529.5	447.0	-3728.6	2839.5	-545.0	402.7	-3127.5	1512.3
PAIRCNDD	12638.9	7373.9	1878.7	59671.0	12650.0	7446.7	2837.7	59671.0
PDECKD	8142.4	5790.7	1134.7	61195.4	7750.5	4254.6	1959.8	48523.4
PFIREPLD	9510.4	5071.5	3200.0	53267.6	9082.2	4193.8	3200.0	49622.9
PSDRANK	-309.0	729.5	-5474.8	6282.8	-207.5	476.4	-2274.5	1858.3
PBEACH	-631.9	4077.6	-251114	57555.2	-352.1	1124.2	-23057.7	5777.8
PFECAL	3.8	76.6	-944.3	1036.8	9.6	87.2	-944.3	1036.8
PSECCHI	33.9	41.9	1.6	645.1	43.0	47.6	5.2	645.1
ADJINC	15626.0	25216.2	-241102	280592.2	12043.0	18513.9	-115500	202278.4
	10665				7193			

Table 6.14 Estimated Marginal Implicit Prices for Secchi COR and Secchi SIG Data: CBG Case

6.4 Conclusion

The first stage of hedonic price estimation is conducted in this chapter and the marginal implicit prices are computed based on the estimated results for each observation for both the individual houses and the census block group cases. Estimated results are mixed for fecal coliform. In six clusters, the fecal coefficients are estimated with negative signs for the case of individual houses while it is negative for five clusters for the census block group case. As for the secchi readings, most of the clusters are estimated with the expected sign. Average MIPs are computed for three types of data set. The first set includes all 10655 observations. The second set consists of the houses whose sales prices are affected negatively (positively) by fecal (secchi) regardless of the significance (COR Data). the third set is composed of the houses whose sales prices are negative (positive) and significantly influenced by fecal (secchi) (SIG Data).

Computed MIPs are -21.6 dollars for fecal with the COR data and -30.5 dollars for the SIG data, indicating a marginal increase in the fecal coliform will decrease the housing price by 21.6 dollars for houses in the COR data and by 30.5 dollars in the SIG data. As for water clarity, we found that an increase in the water clarity will increase the housing price by 40.5 dollars for the COR data and 56 dollars for the SIG data in the case of individual house. Computed MIPs for the case of census block group is -18 dollars for fecal COR data and -53.6 dollars for fecal SIG data. They are 33.9 dollars for secchi COR data and 43 dollars for secchi SIG data. The differences arise from the different definitions of cluster boundaries.

CHAPTER 7

THE SECOND STAGE OF THE HEDONIC STUDY ON LAKE ERIE WATER QUALITY

7.1 The Model

In the second stage of hedonic price analysis, demand for water quality is estimated by using fecal and secchi variable. Clusters determined in the cluster analysis and following merging based on chow test are considered forming housing submarket in the region. Therefore, we have ten submarkets for both individual houses and CBG case. Marginal implicit prices (MIP) estimated and identified submarkets are used to estimate demand functions for both water quality measurements.

The structure of the model is

$$WQ = f(P_{WQ,}\, \mathbf{P_S},\, \mathbf{P_C}, \mathbf{Z})$$

where WQ is quantity of water quality variable (fecal or secchi), $P_{WQ,}$ is marginal price of water quality measurements (fecal or secchi), $\mathbf{P_S}$, is a price vector of substitutes to water quality, $\mathbf{P_C}$ is a price vector of complements to water quality, and $\mathbf{Z}$ is a vector of demographic characteristics. The model specification used for estimation is as follows.

$$Fecal = \beta_1 PFecal + \beta_2 PLOTACR + \beta_3 PBATHN + \beta_6 PDECK$$
$$+\beta_9 PBEACH + \beta_{10} PSECCHI$$
$$+\beta_{11} PCTBACH + \beta_{12} PCT0_{17} + \beta_{13} PCTBLACK + \beta_{14} PCTSINGLE$$
$$+\beta_{15} ADJINC$$

$$Secchi = \beta_1 PSECCHI + \beta_2 PLOTACR + \beta_3 PBATHN + \beta_6 PDECK$$
$$+\beta_9 PBEACH + \beta_{10} PFECAL$$
$$+\beta_{11} PCTBACH + \beta_{12} PCT0_{17} + \beta_{13} PCTBLACK + \beta_{14} PCTSINGLE$$
$$+\beta_{15} ADJINC$$

The own price of water quality is expected to have negative relationship with quantity consumed. Variables of complement to water quality are expected to have negative signs while it is positive for substitutes. Because water quality variables are included in the first stage hedonic price function by taking logarithm, the MIPs of water quality are endogenous in the consumer's choice problem due to the fact that the price of water quality a house owner will face is a function of the quantity of the water quality they choose to consume. Therefore, in the demand function estimation in the second stage, we employ two-stage least squares (2SLS) estimation method in order to handle this endogeneity.

Because the water quality prices are nonlinear, we have to linearize the budget constraint around the chosen consumption bundle (Palmquist (1988), Boyle, Poor and Taylor (1999), Taylor (2000)). Adjusted income is derived as $Y_a = Y - SP + HP$ where

Y is household income, SP is sales price of a house and HP is the hedonic price of a house computed with estimated MIPs and the actual consumption bundle of housing features. In our case, HP is equal to

$$HP = CONSTRANT + \widehat{P_{LOTACR}}^{i} LOTACR^{i} + \widehat{P_{BLDGSF}}^{i} BLDGSF^{i} + \widehat{P_{BATHN}}^{i} BATHN^{i}$$
$$+ \widehat{P_{GRGSQF}}^{i} GRGSQF^{i} + \widehat{P_{AGE}}^{i} AGE^{i} + \widehat{P_{AIRCNDD}}^{i} AIRCNDD^{i}$$
$$+ \widehat{P_{DECKD}}^{i} DECKD^{i} + \widehat{P_{FIREPLD}}^{i} FIREPLD^{i} + \widehat{P_{SDRANK}}^{i} SDRANK^{i}$$
$$+ \widehat{P_{BEACH}}^{i} BEACH^{i} + \widehat{P_{FECAL}}^{i} FECAL^{i} + \widehat{P_{SECCHI}}^{i} SECCHI^{i}$$

Where $\hat{P}$s are MIPs estimated and it is multiplied by the relevant quantity of the variable home owner i consumes.

Instrumental variables are listed as follows.

- Submarket dummy variables
- Demographic variables including adjusted income
- Interaction terms of submarket dummies and demographic variables
- Marginal implicit prices included in the demand estimation model

7.2 Data

In addition to marginal implicit prices derived from the first stage of the estimation and adjusted income specified in the previous section, we include demographic variables adopted from census block group level census data. Demand function for environmental variable includes age and education (Beron *et.al* (2003)) as the demand shifter (Taylor (2003)). In addition to these two factors, we also include race and marital status factors in

our model as in Palmquist (1984). Four variables are 1. percentage of population with bachelor's degree, 2. percentage of kids population age 0 to 17, 3. percentage of black population and 4. percentage of single population within a census block group.

7.3 Estimated Results: Individual Houses Case

Three forms of demand function (linear, semi log and log log) are estimated for three set of data as specified in previous chapter. Since fecal coliform count is a "bad", not a "good", the expected sign of its own price is negative. In order to estimate non-linear demand function, we multiply MIP of fecal by minus one. In equation, this transformation is expressed as

$$(1) \quad Q = \beta P + C \quad where\ P < 0$$

$$(2)\ Q = -\beta^* P^* + C \quad where\ P^* = -P, \beta = -\beta^* \ .$$

We estimate equation (2) and its estimated coefficient is $-\beta^*$ which is expected to be positive. This transformation does not affect the magnitudes or signs of other coefficients. We take logarithm of *P** for semi log model and both *Q* and *P** for log log model. The semi log model is estimated as $Q = -\beta^* lnP^* + C$ (where $-\beta^*$ is the estimated coefficient), and the form is retransformed back to original definition of β and *P* for the derivation of the inverse demand function as follows.

$$Q = -\beta^* lnP^* + C$$

$$\Longrightarrow \quad lnP^{*} = \frac{Q - C}{(-\beta^{*})} \quad \Rightarrow \quad P^{*} = \exp\left[\frac{Q - C}{(-\beta^{*})}\right] \Rightarrow \; P = -\exp\left[\frac{Q - C}{-(-\beta^{*})}\right]$$

The same kind of transformation is done for log log specification as shown below.

$$lnQ = -\beta^{*} lnP^{*} + C$$

$$\Longrightarrow \quad lnP^{*} = \frac{lnQ - C}{(-\beta^{*})} \quad \Rightarrow \quad P^{*} = \exp\left[\frac{lnQ - C}{(-\beta^{*})}\right] \Rightarrow \; P = -\exp\left[\frac{lnQ - C}{-(-\beta^{*})}\right]$$

7.3.1 Two Stage Least Squares Result

In this section, the estimated demand functions for two water quality variables (fecal and secchi) and three data set (all, COR and SIG) estimated with three functional forms (linear, semi log and log log) are reported. Two stage least squares (2SLS) and Non-linear 2SLS (N2SLS) are conducted by using SAS program. In order to determine which functional form is the best fit to our data, we first plot the water quality variables (fecal and secchi) against their marginal implicit prices. The figures are shown from Figure 7.1 through 7.6. In these figures, the figure in upper left corner is the histogram of MIP and the one in the lower right corner is the histogram of quantity of water quality. Upper right corner figure is the one plotting quantity (x) against MIP (y).

Although it is difficult to say the shape of the function from Figure 7.1 since MIP of fecal include both positive and negative values, the figure implies that the relationship between quantity and the price is non-linear. Figure 7.2 also shows that the quantity of secchi and its MIP are non-linear.

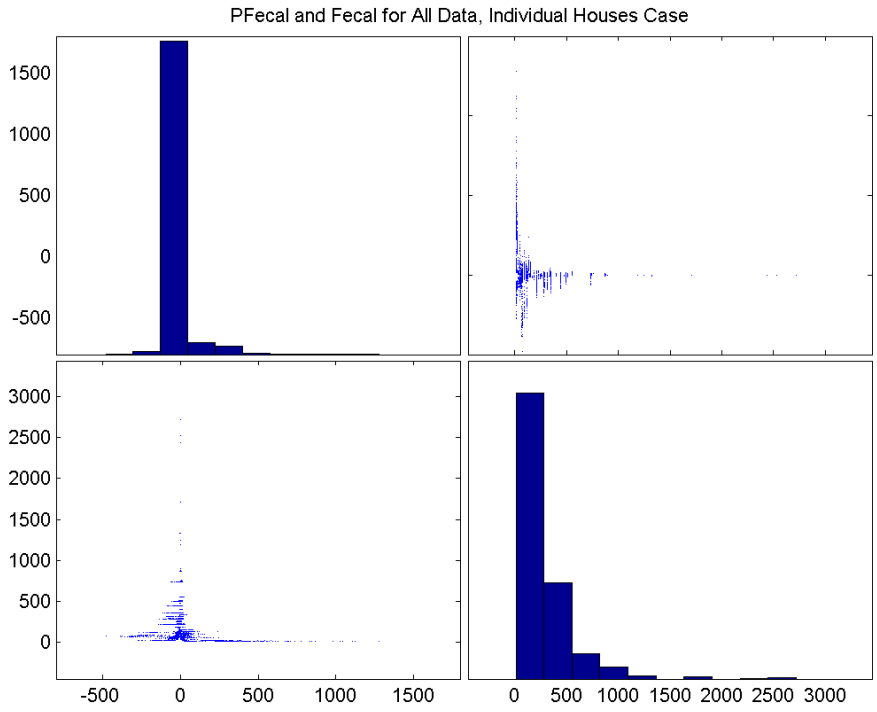


Figure 7.1. Quantity and MIP of Fecal Coliform Counts, All Data: IH Case.

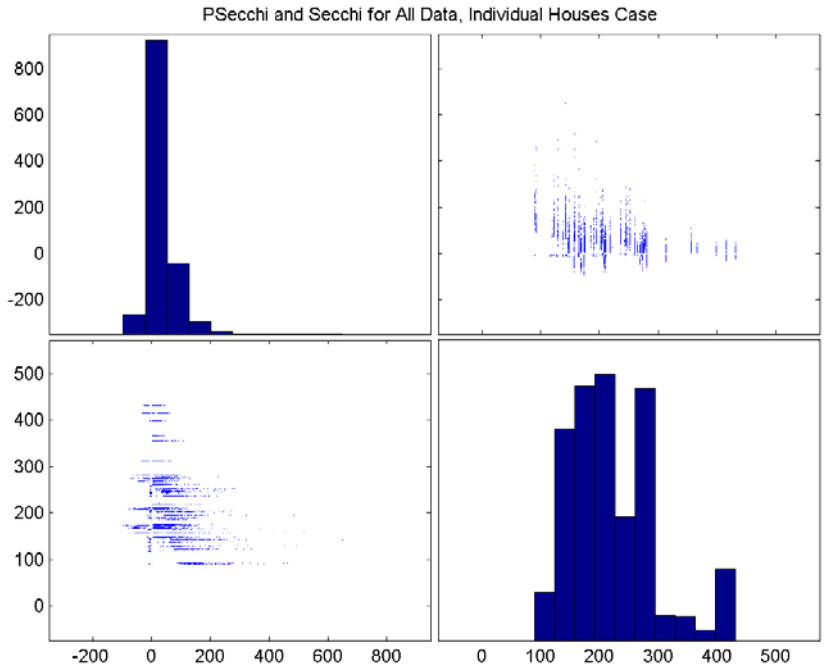


Figure 7.2. Quantity and MIP of Secchi Depth Readings, All Data: IH Case.

Given these observations, we are going to estimate semi log and log log demand equation for COR and SIG data although we cannot do so for ALL data due to negative MIPs. For COR and SIG data for fecal coliform as Figure 7.3 and 7.4 indicate, either semi log or log log function looks more appropriate functional form than linear function. Non-linear relationship is also observed clearly for secchi depth case for all types of data. Although we estimate linear specification for comparison reason, the figures imply that non-linear (semi log or log log) equations are more appropriate for our data.

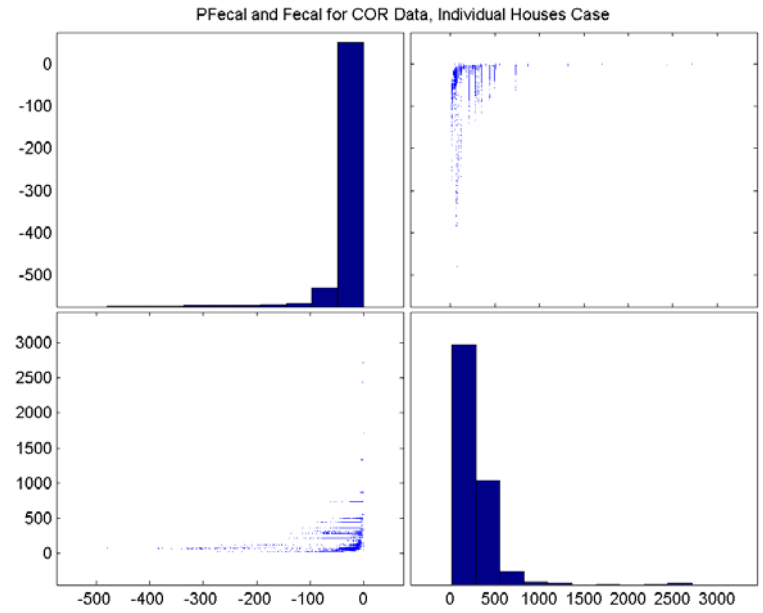


Figure 7.3. Quantity and MIP for Fecal Coliform Counts, COR Data: IH Case

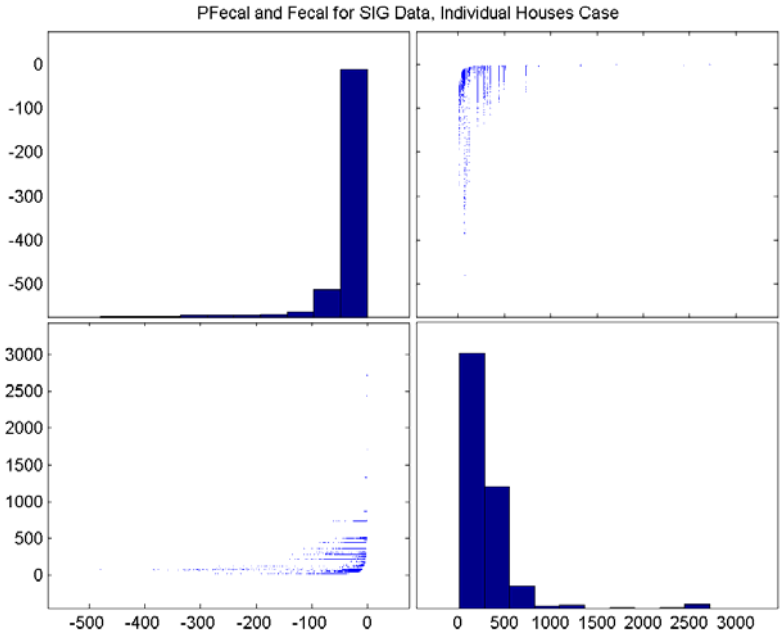


Figure 7.4. Quantity and MIP for Fecal Coliform Counts, SIG Data: IH Case

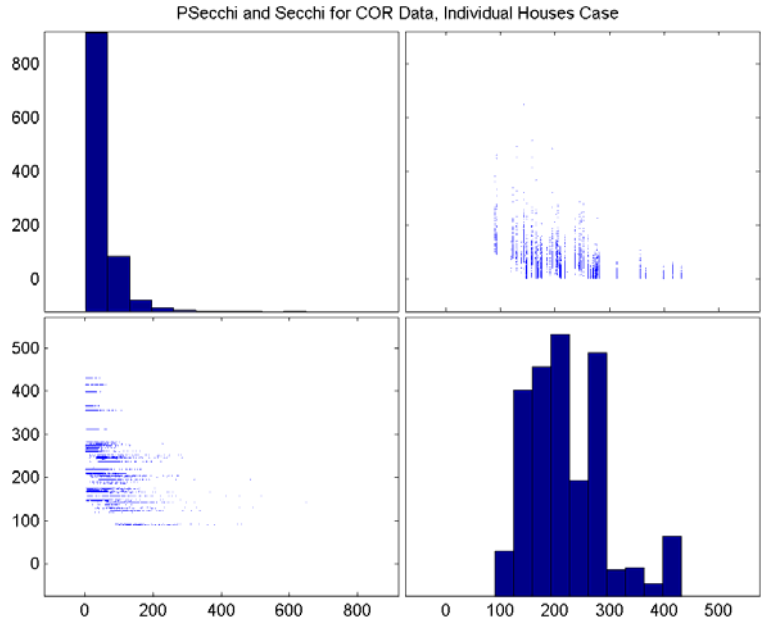


Figure 7.5. Quantity and MIP for Secchi Depth Readings, COR Data: IH Case Figure

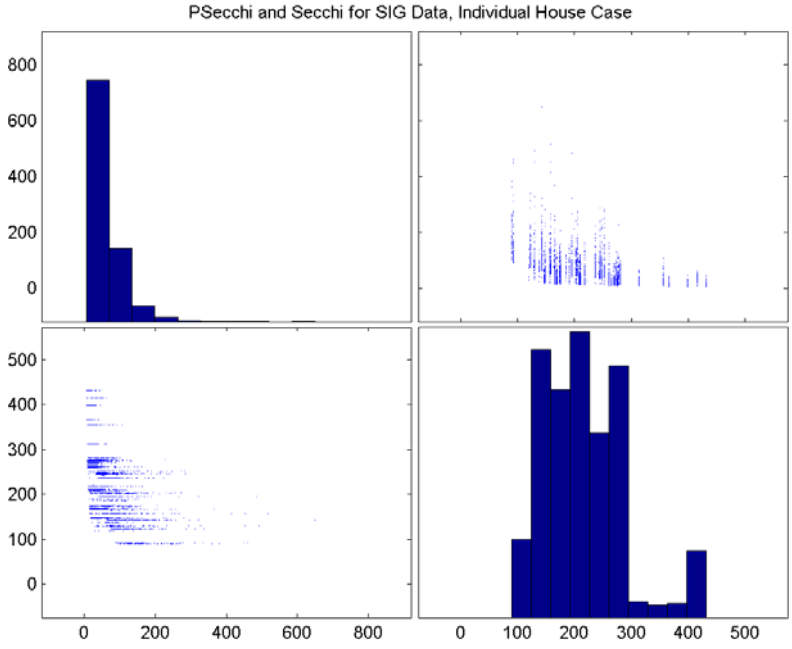


Figure 7.6. Quantity and MIP for Secchi Depth Readings, SIG Data: IH Case

We first estimate linear model with all data for both fecal and secchi cases. Marginal implicit price (MIP) of fecal is expected to be negative for the fecal demand equation as its own price. Since fecal coliform is a "bad", the influence of other prices on the demand for fecal is considered as follows. Note that the quantity demanded is expressed with lower case, q and demand is in upper case, Q.

$$P_o \uparrow \Rightarrow\ q_o \downarrow \Rightarrow \begin{matrix} bad & \uparrow complement \\ Q_{fecal} \downarrow & substitute \end{matrix}$$

where P_o is the price of goods o, Q_o is the quantity demanded for goods o, and Q_{fecal} is the demand for fecal. An increase in the price of normal goods will decrease the quantity demanded of the good, and it will increase the demand for fecal if they are complements and it will decrease fecal if they are substitutes. Therefore, the estimated sign of complements is positive while it is negative for substitutes.

As for secchi demand equation, substitutes will have positive signs and complements will have negative signs as shown below.

$$P_o \uparrow \Rightarrow \begin{matrix} good \\ q_o \end{matrix} \downarrow \Rightarrow \begin{matrix} good & \uparrow & substitute \\ Q_{secchi} \downarrow & & complement \end{matrix}$$

$$P_o \uparrow \Rightarrow \begin{matrix} bad \\ q_o \end{matrix} \uparrow \Rightarrow \begin{matrix} good & \uparrow & substitute \\ Q_{secchi} \downarrow & & complement \end{matrix}$$

It is because an increase in the price of good decrease the quantity demanded of the good and it will then increase the demand for secchi if they are substitute and decrease the demand if they are complements. It is same for MIP for bads since an increase in the

price of reducing a bad will increase the quantity demanded of the bad, and if it has substitutes relationship with secchi, the sign will be positive while it is negative if complements.

The sign for MIP of secchi will be positive if it is a complement to fecal and it is negative if it is a substitute. If water clarity is considered to be a complement to fecal coliform counts, the sign will be negative and if it is a substitute, the sign will be positive.

We expect percentage of bachelor's degree to be negative for fecal and positive for secchi, percentage of children between 0 and 17 to be negative for fecal and positive for secchi, percentage of black to be positive for fecal and negative for secchi, adjusted income to be negative for fecal and positive for secchi. There is no *a priori* expectation for the signs for the percentage of single variable. The estimated results of 2SLS is shown in Table 7.1 for both fecal and secchi outcome for linear model by using all data.

The own price of fecal coliform is estimated as positive. This is mainly because estimated marginal implicit price for fecal includes both positive and negative values. Secchi depth reading is estimated as substitute for fecal coliform while distance to the closest beach is not statistically different from zero. Demographic variables (percentage of children under 18, percentage of black population, and adjusted income) also have opposite signs from the existing studies have suggested. Our results indicate that the higher the percentage of children, the higher the demand for fecal coliform on the beaches, and the higher the rate of the black population, the lower the demand for fecal coliform counts. The adjusted income is estimated as positive and significant, implying that the higher the adjusted income of the household is, the higher the demand for fecal

coliform is. Although these are unexpected results, the positive influence of percentage of children under 17 and the negative effect of percentage of black population are consistent outcome throughout all model specifications and data types.

As for secchi case with all data, we found negative own price effect on the quantity demanded, and fecal coliform is a substitute of waster clarity as we found in fecal demand estimation. The distance to the beach is not estimated as significantly influencing the demand in this case. PCTBACH, PCTSINGLE and ADJINC have positive relationship with the demand for water clarity, while we found that PCTBLACK is negatively related with the demand for water clarity. As oppose to the results from the fecal demand equation, these are expected outcomes. PCT0_17 is not statistically significant in this case.

The estimated results for fecal with COR data and all three functional forms are shown in Table 7.2. The own price of fecal multiplied by minus one is estimated negative for linear model while it is positive for non-linear models. Since we excluded observations with positive MIP of fecal when we formed COR data set, this positive and significant outcomes are somewhat surprising. If we consider this outcome with the results from the first stage estimation, it is possible to conclude that the influence of the possible omitted variable which is related to fecal coliform variable persist both all and COR observations. The distance to the closest beach is not statistically significant for all three specifications. Water clarity is revealed to be a substitute to fecal coliform and this result is consistent for all specifications for COR data. We found that PCT0_17 is positive significant for all functional forms while PCTBLACK is negative significant.

Adjusted income does not have significant influence on the fecal quantity demanded in all specifications. The results for SIG data set is listed in Table 7.3. For this set of data, the own price of fecal is all negative for all specifications as we expect. PSECCHI is not significant for the linear and semi log form while it is negative significant for the log log functional form indicating that water clarity is a substitute to fecal variable as we observed in all and COR data outcomes. The distance to the beach is not significant for any case. PCT0_17 and PCTSINGLE are positively influencing the fecal demand while PCTBLACK is negatively affecting the demand. ADJINC is not significant for linear and semi log models while it is negative and significantly for log log specification. By excluding the insignificant results from the observations from the first stage estimation, we pick up the "bad" aspect of fecal coliform on the beaches for this set of data more explicitly.

Fecal All Linear			Secchi All Linear		
CONSTANT	157.53	***	CONSTANT	194.99	***
	(5.86)			(29.15)	
PFECAL*	0.80	***	PSECCHI	-0.33	***
	(13.24)			(-18.32)	
PLOTACR	-0.55	***	PLOTACR	0.04	
	(-5.61)			(1.54)	
PBATH	0.002	***	PBATH	-0.0001	
	(2.93)			(-0.23)	
PDECK	-0.004	***	PDECK	0.002	***
	(-5.45)			(11.71)	
PBEACH	0.0001		PBEACH	-0.00001	
	(0.39)			(-0.21)	
PSECCHI	-0.50	***	PFECAL	0.03	***
	(-7.81)			(3.43)	
PCTBACH	0.30		PCTBACH	0.52	***
	(0.41)			(2.89)	
PCT0_17	5.29	***	PCT0_17	-0.01	
	(7.61)			(-0.04)	
PCTBLACK	-3.22	***	PCTBLACK	-0.21	**
	(-8.01)			(-2.07)	
PCTSINGLE	1.15		PCTSINGLE	0.92	***
	(1.44)			(4.63)	
ADJINC	0.0007	***	ADJINC	0.00010	***
	(4.98)			(2.99)	
AdjR2	0.03			0.11	
N	10655.00			10655.00	

Table 7.1. 2SLS Estimated Result for Fecal and Secchi with All Data: Individual Houses Case

	Fecal COR					
	Linear		Semilog		Loglog	
CONSTANT	-13.77		-10.61		4.04	***
	(-0.30)		(-0.21)		(26.90)	
PFECAL*	-1.49	***	21.92	***	0.04	**
	(-5.70)		(5.48)		(2.95)	
PLOTACR	0.06		0.02		0.001	***
	(0.56)		(0.20)		(2.74)	
PBATH	0.004	***	-0.01	***	-0.00001	***
	(3.64)		(-3.94)		(-2.73)	
PDECK	0.003	**	0.01	***	0.00002	***
	(2.34)		(3.65)		(3.65)	
PBEACH	0.0001		-0.0001		-0.000001	
	(0.41)		(-0.64)		(-1.41)	
PSECCHI	-0.47	***	-0.57	***	-0.01	***
	(-6.29)		(-6.98)		(-25.51)	
PCTBACH	-0.09		-2.50	**	0.003	
	(-0.10)		(-2.39)		(-0.84)	
PCT0_17	9.51	***	8.21	***	0.03	***
	(10.86)		(8.35)		(10.61)	
PCTBLACK	-5.61	***	-6.60	***	-0.02	***
	(-9.31)		(-10.12)		(-12.35)	
PCTSINGLE	1.68		5.83	***	0.03	***
	(0.83)		(2.64)		(4.16)	
ADJINC	0.0003		0.0002		-0.000001	
	(1.54)		(0.69)		(-0.76)	
AdjR2	0.13		-0.02		0.17	
N	5328.00		5328.00		5328.00	

Table 7.2. 2SLS Estimated Result for Fecal with COR Data: Individual Houses Case

	Fecal SIG					
	Linear		Semilog		Loglog	
CONSTANT	-31.13		342.11	***	5.06	***
	(-0.38)		(3.96)		(25.82)	
PFECAL*	-2.72	***	-143.16	***	-0.48	***
	(-9.40)		(-8.60)		(-12.79)	
PLOTACR	-0.06		-0.14		0.001	***
	(-0.40)		(-1.23)		(3.08)	
PBATH	-0.030	***	-0.03	***	-0.00006	***
	(-8.07)		(-10.90)		(-9.24)	
PDECK	0.042	***	0.04	***	0.00011	***
	(10.71)		(13.25)		(14.10)	
PBEACH	-0.0001		0.00003		0.0000002	
	(-0.35)		(0.18)		(-0.46)	
PSECCHI	0.09		0.12		-0.005	***
	(0.78)		(1.35)		(-22.61)	
PCTBACH	-4.31	***	2.23	**	0.01	***
	(-3.70)		(2.04)		(5.97)	
PCT0_17	8.72	***	6.93	***	0.03	***
	(5.70)		(5.33)		(9.60)	
PCTBLACK	-10.03	***	-6.59	***	-0.02	***
	(-10.63)		(-7.21)		(-11.20)	
PCTSINGLE	10.25	***	5.35	**	0.02	***
	(3.25)		(1.97)		(3.97)	
ADJINC	0.0004		0.0001		-0.000001	**
	(1.56)		(0.54)		(-2.00)	
AdjR2	0.19		0.49		0.74	
N	3555.00		3555.00		3555.00	

Table 7.3. 2SLS Estimated Result for Fecal with SIG Data: Individual Houses Case

Estimated two stage least square results for secchi disk depth readings for COR and SIG data set are listed in Figure 7.4 and 7.5, respectively. Unlike fecal demand case, the estimated results are fairly consistent throughout COR and SIG data set and different functional forms, except for adjusted income. Own prices are estimated negative and significant for all cases. As in the case with fecal, fecal coliform is estimated as a substitute for water clarity. The distance to the beach variables are not significant for any case. As for the demographic variables, PCTBACH, PCT0_17, PCTSINGLE have positive significant impacts on the demand for water clarity while it is negative significant for PCTBLACK for all specifications and data sets. This indicates that the households with the higher education level, with more children and/or who are single have higher demand for water clarity.

The only difference in the direction of the influence in demand between COR and SIG data is the signs for the adjusted income. The expected sign of ADJINC is positive in water clarity case as we can observe in COR data set. However, it is negative and significant for linear and log log specification for SIG data while it is not significant for semi log form. As we described earlier, the adjusted income variable is composed of median household income, discounted housing sales price, the estimated marginal implicit prices for other variables included in the first stage hedonic price estimation and their corresponding quantities consumed. Among the five clusters included in SIG data set, four of them have below the overall average median household income and three of them have below the average discounted house sales price.

Therefore, on average, median household income and the house sale price for SIG data are below the entire average of each variable. Therefore, it may possible to state that this result may be due to the differences in income structures for the observations included in each data set.

	Secchi COR					
	Linear		Semilog		Loglog	
CONSTANT	173.09	***	219.73	***	5.37	***
	(25.56)		(31.31)		(178.94)	
PSECCHI	-0.69	***	-22.81	***	-0.11	***
	(-33.44)		(-29.41)		(-33.41)	
PLOTACR	-0.06	**	-0.13	***	-0.001	***
	(-2.41)		(-4.95)		(-4.74)	
PBATH	0.004	***	0.00	***	0.00001	***
	(16.31)		(6.60)		(7.02)	
PDECK	0.003	***	0.00	***	0.00002	***
	(16.36)		(20.34)		(22.65)	
PBEACH	0.0000		0.0000		0.000000	
	(-0.22)		(-0.07)		(-0.32)	
PFECAL	0.11	***	0.08	***	0.00	***
	(11.63)		(9.18)		(11.57)	
PCTBACH	1.39	***	1.29	***	0.01	***
	(6.94)		(6.42)		(6.30)	
PCT0_17	0.38	**	0.69	***	0.00	***
	(2.20)		(3.91)		(3.28)	
PCTBLACK	-0.15		-0.31	***	0.00	***
	(-1.64)		(-3.44)		(-3.72)	
PCTSINGLE	0.60	***	0.79	***	0.00	***
	(3.31)		(4.34)		(4.87)	
ADJINC	0.0002	***	0.0003	***	0.000001	***
	(4.30)		(7.05)		(7.34)	
AdjR2	0.25		0.25		0.29	
N	8754.00		8754.00		8754.00	

Table 7.4. 2SLS Estimated Result for Secchi with COR Data: Individual Houses Case

	Secchi SIG					
	Linear		Semilog		Loglog	
CONSTANT	170.81	***	291.15	***	5.73	***
	(21.82)		(33.87)		(154.47)	
PSECCHI	-0.94	***	-47.48	***	-0.24	***
	(-32.78)		(-32.26)		(-37.71)	
PLOTACR	-0.03		-0.14	***	-0.001	***
	(-1.01)		(-5.13)		(-5.47)	
PBATH	0.008	***	0.01	***	0.00003	***
	(20.71)		(17.85)		(21.50)	
PDECK	0.000		0.00	***	0.00000	
	(1.32)		(2.59)		(1.21)	
PBEACH	0.0000		0.0000		0.000000	
	(0.39)		(0.57)		(0.49)	
PFECAL	0.14	***	0.13	***	0.00	***
	(16.91)		(16.21)		(19.54)	
PCTBACH	1.06	***	1.22	***	0.01	***
	(4.88)		(5.86)		(6.04)	
PCT0_17	0.50	**	0.95	***	0.00	***
	(2.38)		(4.68)		(4.89)	
PCTBLACK	-0.12		-0.32	***	0.00	***
	(-1.28)		(-3.63)		(-4.67)	
PCTSINGLE	0.67	***	1.06	***	0.01	***
	(3.92)		(6.45)		(7.63)	
ADJINC	-0.0001	**	0.0000		0.000000	*
	(-2.47)		(-0.37)		(-1.86)	
AdjR2	0.40		0.45		0.51	
N	5796.00		5796.00		5796.00	

Table 7.5 2SLS Estimated Result for Secchi with SIG Data: Individual Houses Case

7.3.2 Estimated Demand Function: Individual Houses Case

Given the estimated results from two stage least squares discussed in the previous section, we computed demand functions for each functional form and data set. The demand functions for fecal is reported without retransforming the price. The functions for each type can be found in Table 7.6. Demand functions for Fecal have negative slopes

for [COR, Linear] and all functional forms for SIG case. Considering that fecal is a "bad" and its price has been transformed, negative slope is the expected sign. On the other hand, [ALL, Linear] and [COR, Semi log and Log log] cases have positive slopes. As for [All, Linear] case, we assume that the influence of positive MIP for fecal caused this outcome. For the case of non-linear specifications for COR data, we suspect that the possible omitted variable is causing this sign reversal. Therefore, we do not consider the results for COR is credible outcome.

The price elasticity of demand is -0.52 for [SIG, Semi log] case if we compute it with mean fecal coliform value for SIG case, and -0.48 for [SIG, Log log] case. Therefore, the price elasticity of demand is relatively inelastic for both cases and the output from Log log form has lower elasticity.

The inverse demand functions estimated for fecal coliform are plotted in Figure 7.7 for the linear function for all data, Figure 7.8 for the linear functions from COR data and Figure 7.9 for the semi log functions for COR data, Figure 7.10 for log log function for COR data, Figure 7.11 for the linear function for SIG data and Figure 7.12 for non-linear functions for SIG data. Based on the functions listed in Table 7.6, we retransformed the price of the fecal variable as shown earlier and plotted with the original negative prices. Therefore, most of the fecal demand functions are located in the fourth quadrant. The expected shape of inverse demand function for fecal (a bad) is the one as shown in the case [SIG, Semi log and Log log] because at the condition with high fecal coliform, the willingness to pay for the one unit reduction of fecal is higher than the case with the lower initial fecal coliform amount.

	Data Type	Fun.Form			Intercept	Slope	
Fecal	All	Linear	P	=	160.07	1.25	Q
	COR	Linear	P	=	102.06	-0.67	Q
	COR	Semilog	lnP	=	-5.87	0.05	Q
	COR	Loglog	lnP	=	-117.60	28.34	lnQ
	SIG	Linear	P	=	107.83	-0.37	Q
	SIG	Semilog	lnP	=	4.11	-0.0070	Q
	SIG	Loglog	lnP	=	11.61	-2.07	lnQ
Secchi	All	Linear	P	=	748.80	-3.05	Q
	COR	Linear	P	=	277.09	-1.44	Q
	COR	Semilog	lnP	=	8.95	-0.04	Q
	COR	Loglog	lnP	=	48.17	-9.02	lnQ
	SIG	Linear	P	=	229.16	-1.07	Q
	SIG	Semilog	lnP	=	5.77	-0.02	Q
	SIG	Loglog	lnP	=	23.66	-4.18	lnQ

Table 7.6. Estimated Demand Functions: Individual Houses Case

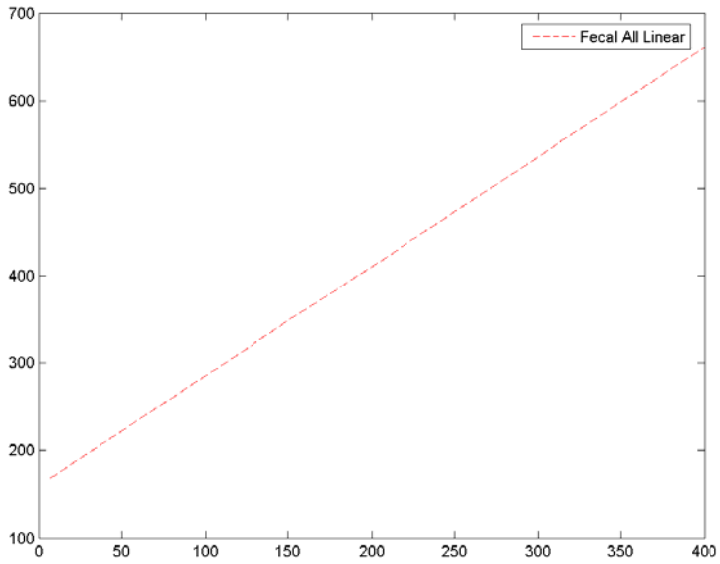


Figure 7.7. Linear Demand Function for Fecal, All Data: IH Case

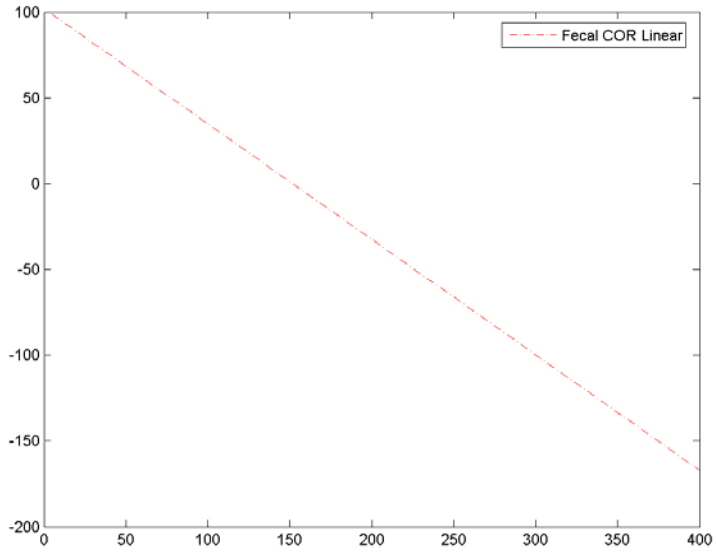


Figure 7.8. Linear Demand Functions for Fecal, COR Data: IH Case

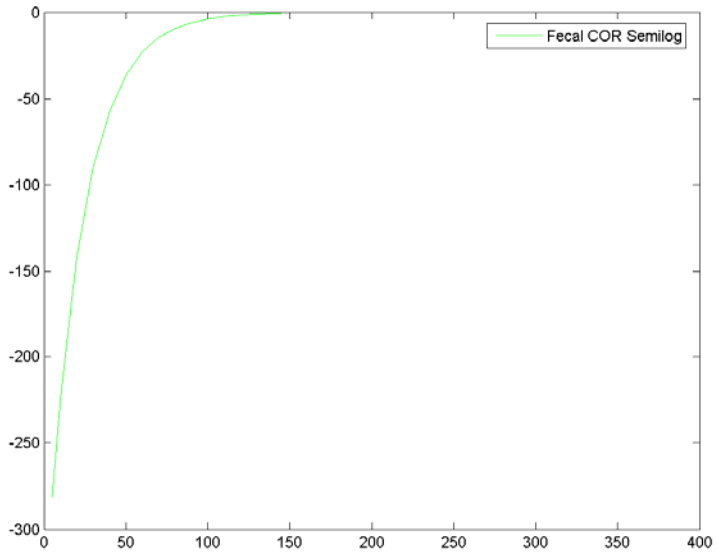


Figure 7.9. Semi log Demand Functions for Fecal, COR Data: IH Case

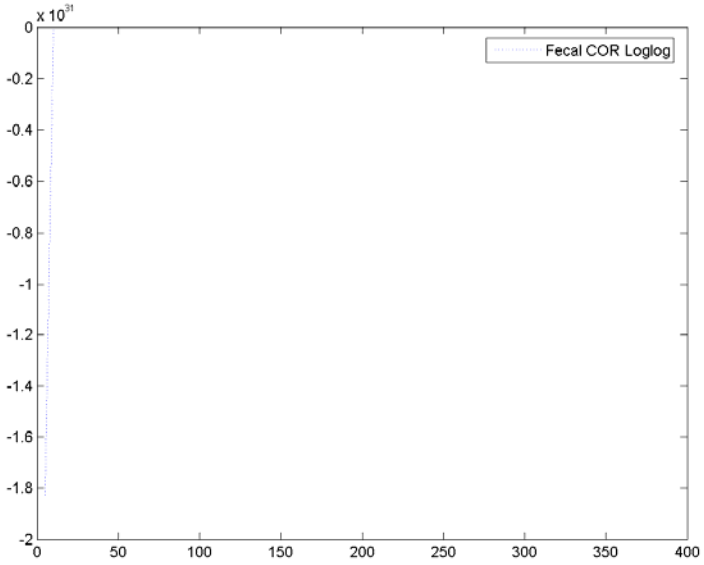


Figure 7.10. Log log Demand Functions for Fecal, COR Data: IH Case

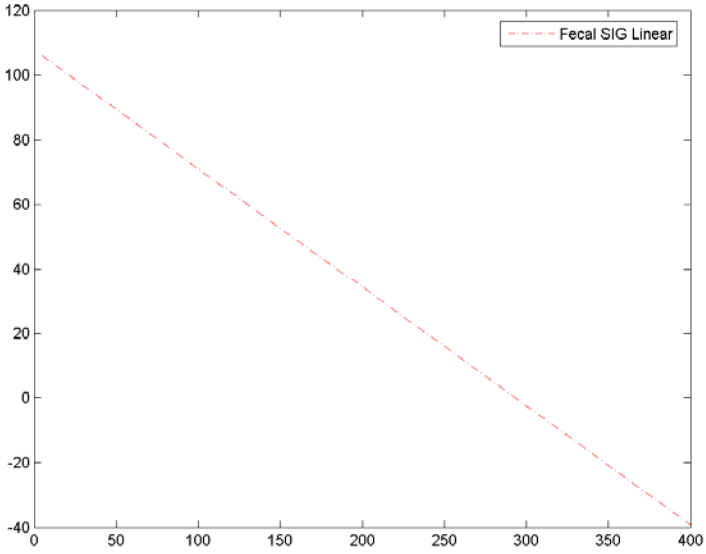


Figure 7.11. Linear Demand Functions for Fecal, SIG Data: IH Case

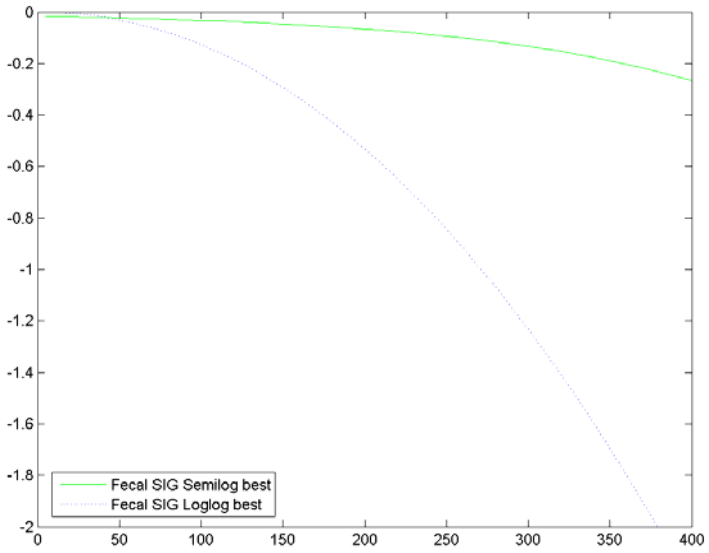


Figure 7.12. Non-linear Demand Functions for Fecal, SIG Data: IH Case

The derived inverse demand function for secchi is also listed in Table 7.6. The price elasticity of demand is -0.10 for [COR, Semi log] case by using the secchi quantity as the average for COR data set, 222 cm while it is -0.11 for [COR, Log log] case. Therefore, derived demand functions indicate very inelastic situation. As for the outcome from SIG data set, the price elasticity of demand computed as -0.22 for the semi log case by using 216 cm as the mean secchi value for this data set, and -0.24 for the log log form. If we compare these with the COR data case, they are less inelastic although these are still very inelastic.

The derived demand functions for secchi readings are shown in Figure 7.13 through 7.15., where Figure 7.13 shows the linear demand function with all data, Figure 7.14. is for the COR dataset case and Figure 7.15. is for the SIG dataset case. The [COR, Semi log] case shows the sleep increase in its slope around 120 cm while it is around 150 cm for the [COR, Log log] case. On the other hand, for the [SIG, Semi log] case, the increase in slope is observed around 100 cm and it is around 120 cm for the [SIG, Log log] case. As the calculation of the price elasticity of demand shows, COR data case has steeper slope. It causes the welfare change for the reduction in water clarity to be extremely high especially for the case that water clarity has been reduced beyond the steep changing point of the slope. Considering the fact that the average water clarity is 222 cm for COR case and 216 cm for SIG case, the linear functions which intersect with x axis around 200 are not reasonable. Therefore, we place more credibility to non-linear demand cases.

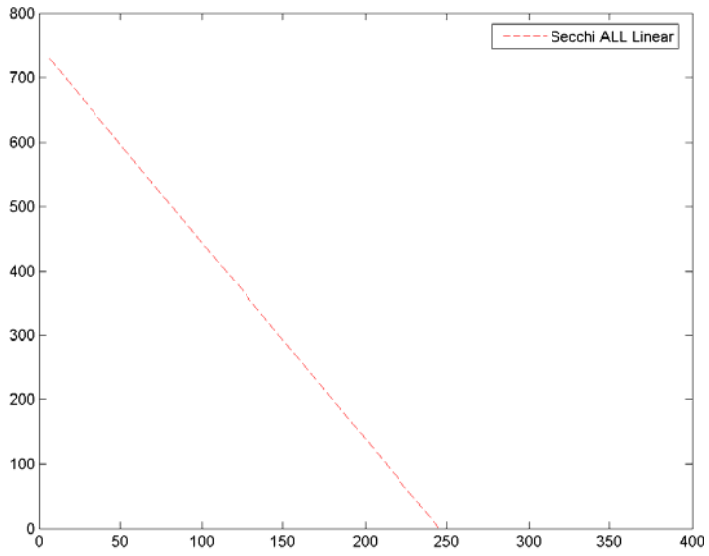


Figure 7.13. Demand Function for Secchi, All Data: IH Case

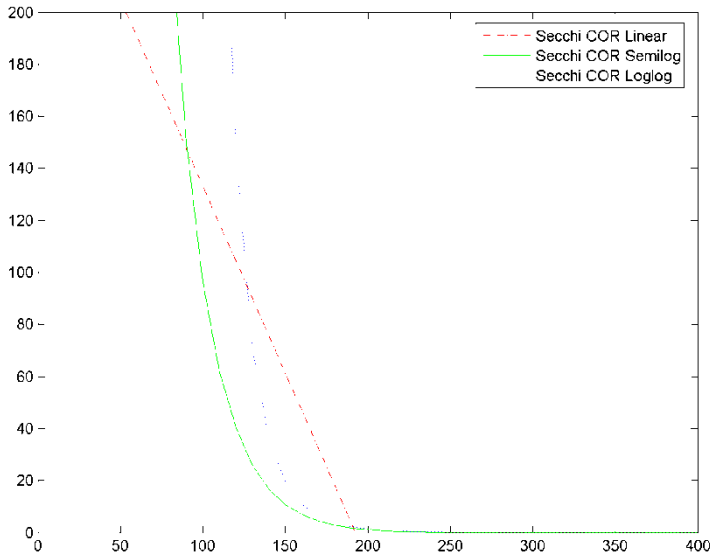


Figure 7.14. Demand Functions for Secchi, COR Data: IH Case

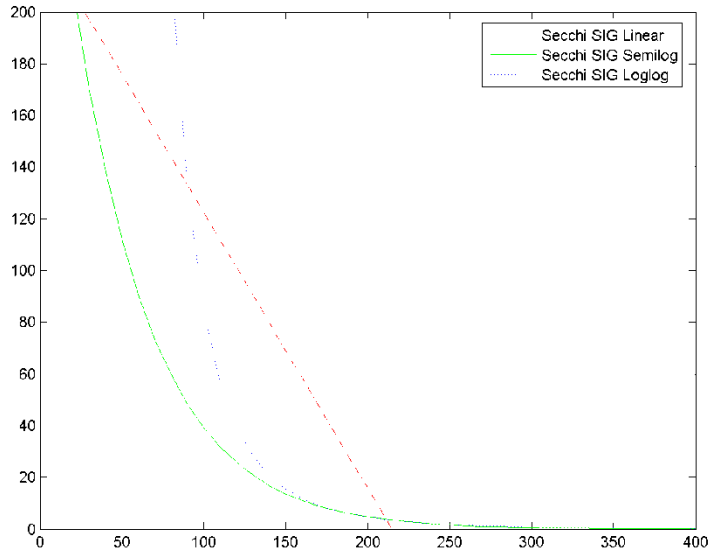


Figure 7.15. Demand Function for Secchi, SIG Data: IH Case

7.3.3 Computed Welfare Change: Individual Houses Case

In this section of the second stage analysis, we are finally able to derive the welfare changes due to non-marginal changes in the water quality variables. As for fecal coliform, we need to retransform the price and its estimated slope coefficients as described in the beginning of section 7.3. By using the retransformed variables, we computed the welfare changes for the changes of the fecal coliform variables to eight different values, four are intending to show the influence for the improvement or the reduction in fecal coliform and the other four is for the degradation or the increase in the counts from the fecal coliform variables' overall mean value, 255 counts per 100 ml.

Although we reported the computed welfare changes for all cases, the computed results for [All, Linear], [COR, Semi log] and [COR, Log log] are based on unreasonable estimated results and difficult to interpret in this setting. We expect the welfare changes to be expressed as negative values since we are handling households' willingness to pay for the reduction of the water quality variable (fecal coliform). Therefore, the [SIG, Linear] case is also not appropriate since the most of the function exist in the first quintile although the slope of the function is negative. The shape of the demand functions estimated for [SIG, nonlinear] cases suggest that welfare changes due to degradation are larger than the changes from improvement for the same amount of the change in quantity, but in different directions. We found that house owners are willing to pay for 20 dollars for the reduction of fecal coliform by 25 counts per 100 ml while it is 24 dollars for the increase in fecal coliform by 25 counts per 100 ml. For the reduction of 150 counts, it is 68 dollars while for the same amount increase in fecal, the welfare change calculated is

230 dollars. Our findings indicate that the fecal coliform counts itself changes the household's welfare by very small amount.

We turn to the discussion for the results of secchi disk depth readings. First of all, because the linear demand functions estimated for all three data set intersect with x axis at around the mean value of secchi variable for each case, the computed welfare measures for the increase in water clarity is estimated as negative. Although it is possible to consider that too high water clarity may cause home owners to have negative willingness to pay because too high water clarity may mean that the lake is experiencing the high level of acidity and the number of fishes living there has been decreasing. However, as we observed the non-linear relationship between the price and the secchi quantity, we place more credibility on the results from non-linear cases. The amount of welfare changes are larger for the houses whose sales prices are influenced significantly by water clarity compared to the COR case. This is expected because we consider the households whose housing values reflect water clarity of the Lake more significantly will have the higher willingness to pay for the water clarity. Their differences are around six to ten times.

The increase in water clarity from 220 cm to 245, 270, 320 and 370 cm will increase the welfare between 8 to 17 dollars for COR case while its decrease causes very dramatic changes in the welfare due to the shape of the demand function estimated. The change for the water clarity reduction below 120 cm has very high changes in welfare especially for Log log case.

As for SIG case, we found that 25 cm changes in water clarity will change the home owner's welfare by 63 dollars and its increase to 176 dollars for the increase by 150 cm. On the other hand, the welfare loss for the decrease in water clarity by 25 cm is 101 dollars while it is 276, 1272 and 8031 dollars for the change by 50, 100 and 150 cm, respectively. Since the Lake water quality is a public good, these welfare change computed is for a house included into the data set. If we want to know the total benefits or damages from water quality changes, we multiply the value we found by the relevant population. For example, if we take the results for SIG data, we multiply the estimated welfare change by the number of houses significantly affected by the water quality in these four counties.

				New Fecal Level (From 255 counts /100 ml)							
				Improvement				Degradation			
			N	230	205	155	105	280	305	355	405
	All	Linear	10655	11600	22417	41700	57850	12383	25550	54233	86050
	COR	Linear	5328	-1528	-2635	-3588	-2858	-1948	-4317	-10316	-17998
	SIG	Linear	3555	466	1163	3245	6246	237	243	-432	-2028
Fecal	COR	Semilog	5328	-0.1	-0.6	-6.5	-64.3	0.0	-0.1	-0.1	-0.1
	SIG	Semilog	3555	-2.2	-4.1	-7.0	-9.0	-2.6	-5.8	-14.0	-25.7
	COR	Loglog	5328	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SIG	Loglog	3555	-19.8	-35.7	-57.2	-68.3	-24.3	-53.6	-128.8	-229.6

Table 7.7. Computed Welfare Change for Fecal (in $ 1996) : Individual Houses Case

				New Secchi Level (From 220 cm)							
				Improvement				Degradation			
			N	245	270	320	370	195	170	120	70
Secchi	All	Linear	10655	1012	120	-7377	-22491	2916	7736	23088	46057
	COR	Linear	8754	-1439	-3777	-11152	-22126	-539	-179	3241	10259
	SIG	Linear	5796	-477	-1621	-5911	-12869	191	1048	4766	11152
	COR	Semilog	8754	8	10	11	11	23	91	904	8188
	SIG	Semilog	5796	61	97	130	142	103	277	1070	3343
	COR	Loglog	8754	10	14	16	17	28	117	2163	164370
	SIG	Loglog	5796	63	104	151	176	101	276	1272	8031

Table 7.8. Computed Welfare Change for Secchi (in $ 1996) : Individual Houses Case

7.4 Estimated Results: Census Block Group Case

In this section, we report the second stage estimation by using the clustering results with census block group. As in the individual houses case, we start by plotting the water quality variables against their marginal implicit prices derived from the first stage of the estimation. Both for the fecal and the secchi cases, the plots suggest non-linear relationship between the prices and the quantities. As for secchi case, COR data coincides with All data indicating all the houses in our sample is affect positively by water clarity although the influence may not be significant.

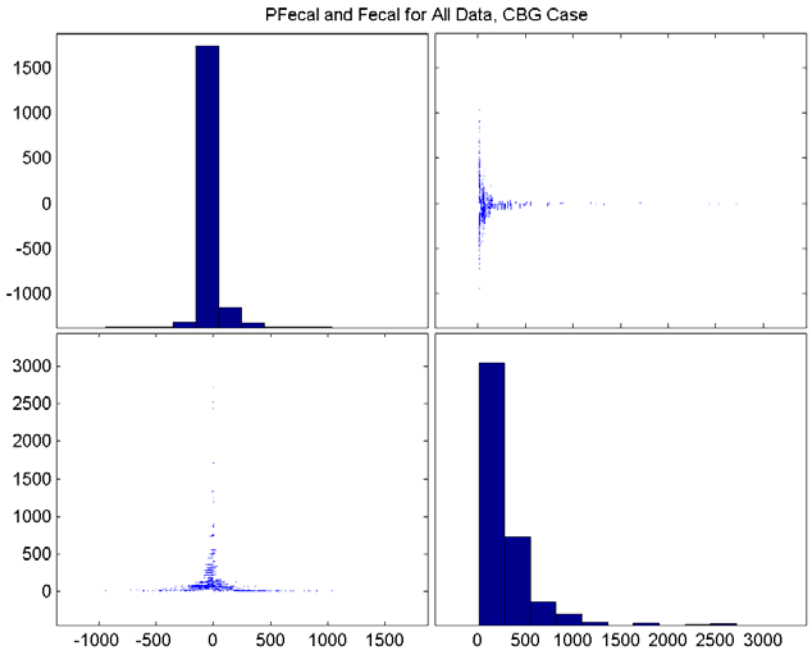


Figure 7.16. Quantity and MIP for Fecal Coliform Counts, All Data: CBG Case

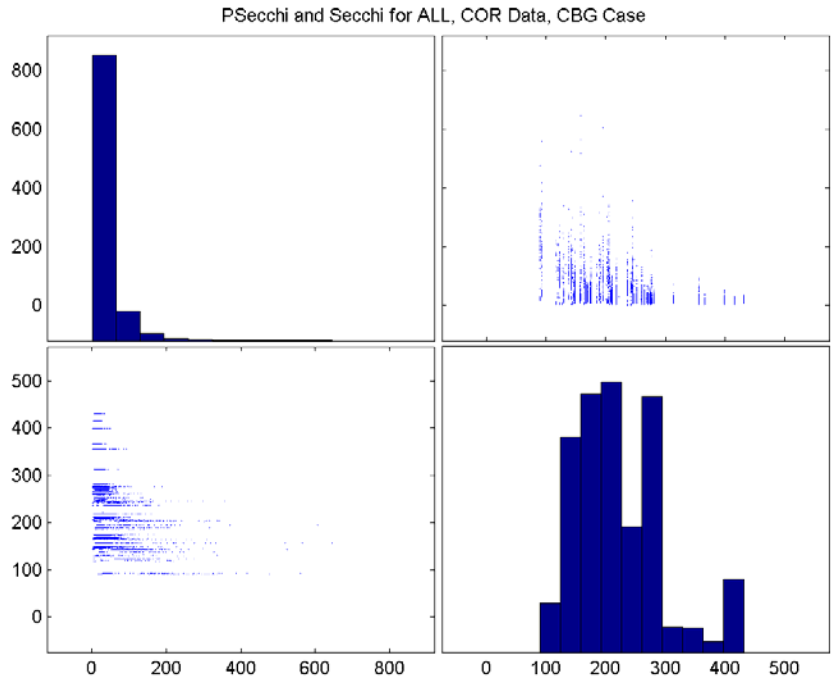


Figure 7.17. Quantity and MIP for Fecal Coliform Counts, All Data: CBG Case

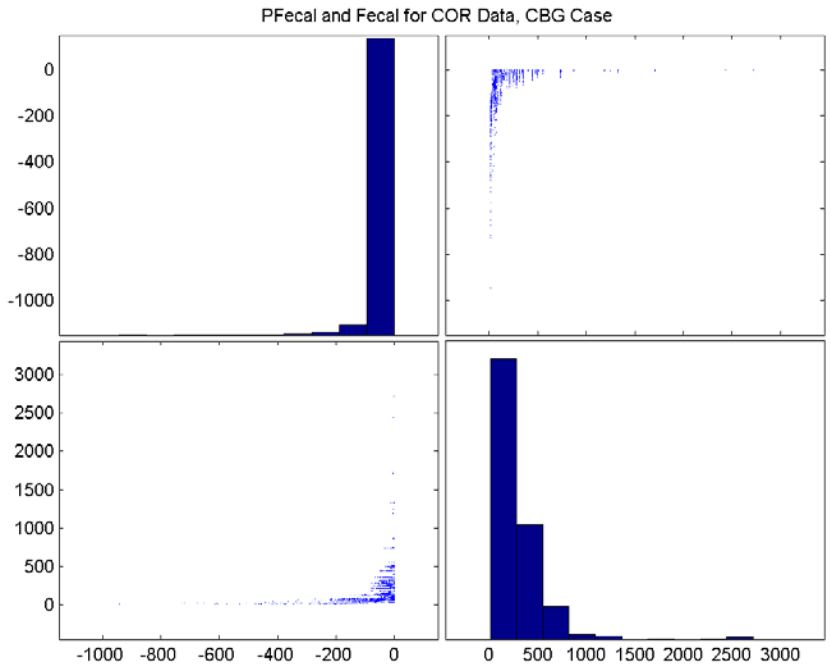


Figure 7.18. Quantity and MIP for Fecal Coliform Counts, COR Data: CBG Case

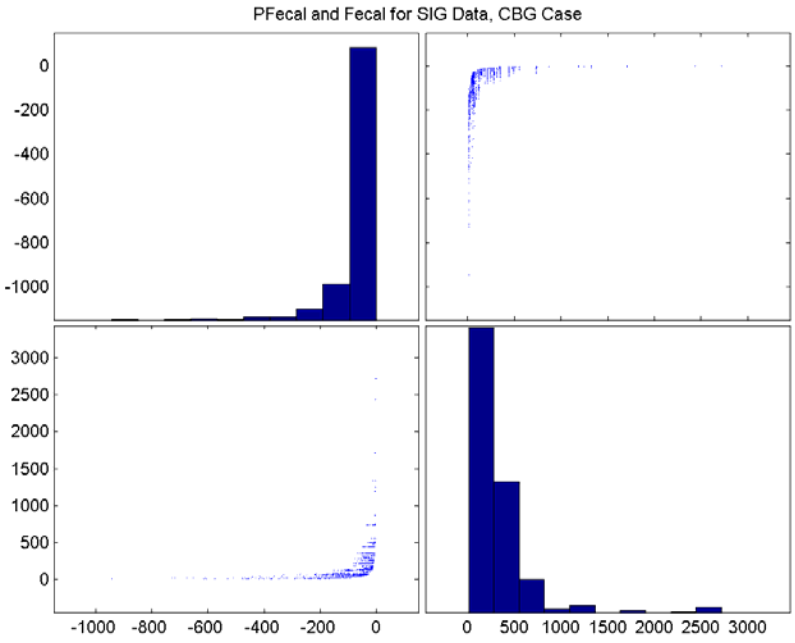


Figure 7.19. Quantity and MIP for Fecal Coliform Counts, SIG Data: CBG Case

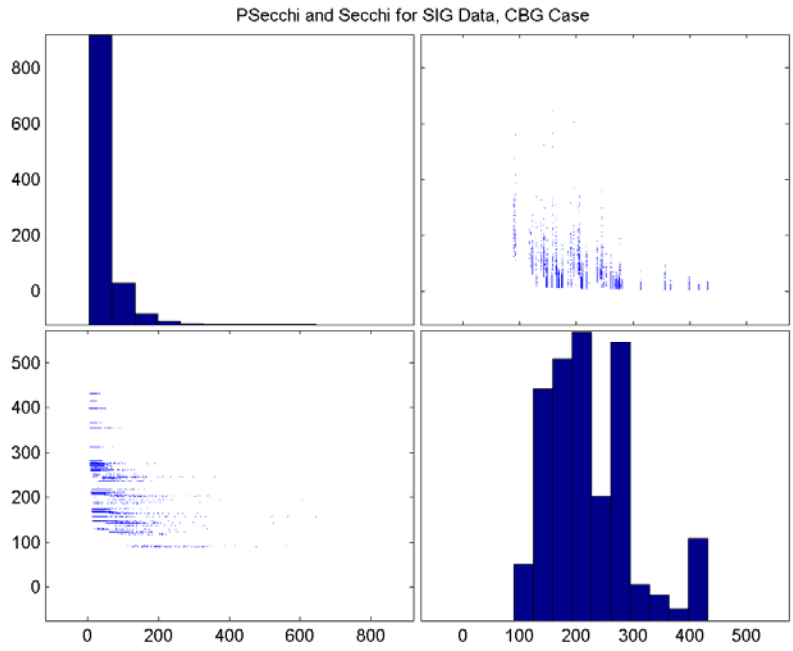


Figure 7.20. Quantity and MIP for Secchi Depth Readings, SIG Data: CBG Case

7.4.1 Two Stage Least Squares Result

The estimated outcomes are listed in Table 7.9 though 7.13. Own prices of water quality are estimated as negative and significant in all cases except for fecal with all data and semi log specification with COR data. The distance to the closest beach is revealed to be a complement to fecal coliform level for all and COR data cases while the effect is not significant for SIG case. Water clarity level is estimated as a substituted for all, [COR, Semi log] and [COR, Log log] case and it is a complement for [COR, Linear], [SIG, Semi log] and [SIG, Log log] case. However, considering the very low level of adjusted R-squares for ALL and COR models (< 0.10), we would like to focus on the results for the SIG data case.

PCTBACH, PCT0_17 and PCTSINGLE are estimated as positive significant while it is negative significant for PCTBLACK case in most of the specifications. These unexpected results coincide with the one from individual houses case. Adjusted income is estimated as positive significant for all and COR cases while it is positive significant for [SIG, Log log] case.

As for the results from secchi demand estimation, we found that in all cases the own price is estimated as negative and statistically significant. The distance to the closest beach is not statistically significant for all or COR data case, but it is negative and significant for SIG dataset indicating that the proximity to the beach is a complement to water clarity. Fecal variable is estimated as positive and significant for most of the cases, indicating that the level of fecal coliform is a substitute to water clarity as we found in the case of individual houses.

PCTBACH is estimated as positive and significant in all cases, implying that the higher educated home owners demand more water clarity. Interestingly, PCT0_17 is not significant for any of the specifications. PCTBLACK has negative significant outcome for COR data while it is positive and significant for SIG data. As oppose to the case with individual houses, the direction of the influence toward water quality is not consistent across different settings and models in the case of census block group. While PCTSINGLE is estimated as positive and significant for most of the cases, ADJINC is positive significant for the [All, Linear] and [COR, Linear] case and not significant for rest of the cases.

Fecal ALL Linear			Secchi ALL Linear		
CONSTANT	140.71	***	CONSTANT	201.91	***
	(5.24)			(30.77)	
PFECAL*	0.75	***	PSECCHI	-0.38	***
	(13.25)			(-19.39)	
PLOTACR	-0.14		PLOTACR	0.04	**
	(-1.52)			(2.05)	
PBATH	0.001		PBATH	-0.001	***
	(0.71)			(-3.44)	
PDECK	0.0003		PDECK	0.003	***
	(0.32)			(13.22)	
PBEACH	0.004	***	PBEACH	0.0002	
	(5.21)			(1.10)	
PSECCHI	-0.98	***	PFECAL	-0.0001	
	(-14.37)			(-0.01)	
PCTBACH	-0.20		PCTBACH	0.71	***
	(-0.27)			(3.97)	
PCT0_17	5.16	***	PCT0_17	-0.18	
	(7.20)			(-1.02)	
PCTBLACK	-3.13	***	PCTBLACK	-0.14	
	(-7.68)			(-1.45)	
PCTSINGLE	1.60	**	PCTSINGLE	0.63	***
	(2.00)			(3.22)	
ADJINC	0.0005	***	ADJINC	0.0001	*
	(3.08)			(1.69)	
AdjR2	0.03			0.14	
N	10655			10655	

Table 7.9. 2SLS Estimated Result for Fecal and Secchi with All Data: CBG Case

	Fecal COR					
	Linear		Semilog		Loglog	
CONSTANT	-122.24	***	-180.80	***	3.66	***
	(-2.90)		(-4.02)		(26.68)	
PFECAL*	-0.93	***	9.66	***	-0.02	***
	(-4.44)		(4.00)		(-2.92)	
PLOTACR	-0.22	*	-0.44	***	-0.0010	***
	(-1.90)		(-3.78)		(-2.90)	
PBATH	0.008	***	0.006	***	0.000020	***
	(7.18)		(5.15)		(6.02)	
PDECK	-0.007	***	-0.006	***	-0.00001	*
	(-6.45)		(-5.13)		(-1.91)	
PBEACH	0.00	**	0.005	***	0.00002	***
	(2.21)		(2.90)		(4.06)	
PSECCHI	0.35	**	-0.46	***	-0.004	***
	(2.17)		(-5.03)		(-14.81)	
PCTBACH	1.82	**	0.83		0.01	**
	(2.09)		(0.94)		(2.20)	
PCT0_17	7.75	***	8.68	***	0.030	***
	(9.43)		(10.06)		(11.38)	
PCTBLACK	-7.61	***	-8.80	***	-0.022	***
	(-10.57)		(-11.41)		(-9.54)	
PCTSINGLE	11.54	***	15.48	***	0.048	***
	(6.64)		(8.36)		(8.51)	
ADJINC	0.0003	*	0.0001		0.0000017	***
	(1.74)		(0.76)		(3.03)	
AdjR2	0.08		-0.001		0.15	
N	7311		7311		7311	

Table 7.10. 2SLS Estimated Result for Fecal with COR Data: CBG Case

	Fecal SIG					
	Linear		Semilog		Loglog	
CONSTANT	-320.06	**	553.98	***	7.84	***
	(-2.10)		(3.88)		(65.07)	
PFECAL*	-2.31	***	-212.96	***	-1.12	***
	(-5.20)		(-11.62)		(-72.35)	
PLOTACR	0.04		-0.03		-0.0001	
	(0.21)		(-0.20)		(-0.83)	
PBATH	-0.047	***	-0.014		-0.000100	***
	(-3.28)		(-1.52)		(-12.58)	
PDECK	0.062	***	0.027	**	0.00019	***
	(3.34)		(2.32)		(19.33)	
PBEACH	0.00		0.000		0.00000	
	(-0.63)		(0.06)		(-0.91)	
PSECCHI	-0.18		0.55	***	0.001	***
	(-0.64)		(2.58)		(4.36)	
PCTBACH	4.31	**	-0.82		0.01	***
	(2.33)		(-0.65)		(10.15)	
PCT0_17	7.82	***	3.36	*	0.005	***
	(2.92)		(1.75)		(3.20)	
PCTBLACK	-23.46	***	-10.75	***	0.005	***
	(-8.13)		(-4.95)		(2.91)	
PCTSINGLE	29.44	***	14.83	***	-0.011	***
	(5.14)		(3.43)		(-2.93)	
ADJINC	0.0001		-0.0001		-0.0000017	***
	(0.17)		(-0.27)		(-7.63)	
AdjR2	0.17		0.59		0.97	
N	2275		2275		2275	

Table 7.11. 2SLS Estimated Result for Fecal with SIG Data: CBG Case

	Secchi COR					
	Linear		Semilog		Loglog	
CONSTANT	201.91	***	214.46	***	5.37	***
	(30.77)		(30.69)		(179.64)	
PSECCHI	-0.38	***	-12.61	***	-0.06	***
	(-19.39)		(-14.97)		(-17.82)	
PLOTACR	0.04	**	0.05	**	0.0002	**
	(2.05)		(2.25)		(2.15)	
PBATH	-0.001	***	-0.001	***	-0.000004	***
	(-3.44)		(-4.43)		(-4.21)	
PDECK	0.003	***	0.003	***	0.00002	***
	(13.22)		(14.47)		(15.17)	
PBEACH	0.00		0.000		0.00000	
	(1.10)		(1.21)		(0.92)	
PFECAL	0.00		0.03	***	0.000	***
	(-0.01)		(3.76)		(4.94)	
PCTBACH	0.71	***	0.88	***	0.00	***
	(3.97)		(4.93)		(4.20)	
PCT0_17	-0.18		0.09		-0.001	
	(-1.02)		(0.52)		(-0.79)	
PCTBLACK	-0.14		-0.22	**	-0.001	**
	(-1.45)		(-2.22)		(-2.55)	
PCTSINGLE	0.63	***	0.86	***	0.004	***
	(3.22)		(4.42)		(4.96)	
ADJINC	0.0001	*	0.0000		0.0000002	
	(1.69)		(1.32)		(1.57)	
AdjR2	0.14		0.15		0.17	
N	10655		10655		10655	

Table 7.12. 2SLS Estimated Result for Secchi with COR Data: CBG Case

	Secchi SIG					
	Linear		Semilog		Loglog	
CONSTANT	190.47	***	259.41	***	5.58	***
	(20.85)		(26.74)		(137.93)	
PSECCHI	-0.50	***	-27.12	***	-0.13	***
	(-20.30)		(-22.14)		(-25.95)	
PLOTACR	0.03		0.06		0.0003	*
	(0.65)		(1.55)		(1.75)	
PBATH	0.002	***	0.001	***	0.000005	***
	(4.66)		(2.61)		(2.77)	
PDECK	0.003	***	0.003	***	0.00002	***
	(8.51)		(11.95)		(13.08)	
PBEACH	-0.01	***	-0.004	***	-0.00003	***
	(-6.70)		(-4.92)		(-7.65)	
PFECAL	0.06	***	0.10	***	0.001	***
	(5.33)		(9.51)		(12.41)	
PCTBACH	1.19	***	1.30	***	0.01	***
	(4.78)		(5.55)		(5.14)	
PCT0_17	0.11		0.06		-0.001	
	(0.45)		(0.27)		(-0.77)	
PCTBLACK	0.24	*	0.17		0.001	**
	(1.94)		(1.43)		(2.33)	
PCTSINGLE	0.22		0.40	**	0.002	**
	(1.07)		(2.05)		(2.23)	
ADJINC	-0.0001		-0.0001		-0.0000003	
	(-0.89)		(-1.52)		(-1.61)	
AdjR2	0.22		0.29		0.34	
N	7193		7193		7193	

Table 7.13. 2SLS Estimated Result for Secchi with SIG Data: CBG Case

7.4.2 Estimated Demand Function: CBG Case

Estimated results of two stage least squares are used to calculate inverse demand functions. The derived functions are listed in Table 7.14. The slopes of fecal coliform demand functions are mostly negative except for the ALL case and [COR, Semi log]

case. The price elasticity of demand for SIG cases are -0.73 for semi log case and -1.12 for log log case. Therefore, we found that especially for the case of log log demand function, the demand is elastic. Even for the semi log case, it is less inelastic comparing to the individual houses case.

As for the secchi case, the slope is negative for all cases. As oppose to the outcome from fecal coliform, the computed price elasticity of demand is lower than the individual houses case. They are -0.06 for COR case and -0.13 for SIG case for both semi log and log log settings. Therefore, we found that the demand for secchi is twice more inelastic for census block group cases.

	Data Type	Fun.Form			Intercept	Slope	
Fecal	All	Linear	P	=	6.70	1.34	Q
	COR	Linear	P	=	220.72	-1.08	Q
	COR	Semilog	lnP	=	12.00	0.10	Q
	COR	Loglog	lnP	=	167.57	-46.40	lnQ
	SIG	Linear	P	=	154.22	-0.43	Q
	SIG	Semilog	lnP	=	4.47	-0.0047	Q
	SIG	Loglog	lnP	=	7.16	-0.89	lnQ
Secchi	All	Linear	P	=	635.92	-2.64	Q
	COR	Linear	P	=	635.92	-2.64	Q
	COR	Semilog	lnP	=	21.94	-0.08	Q
	COR	Loglog	lnP	=	87.53	-15.59	lnQ
	SIG	Linear	P	=	481.24	-2.01	Q
	SIG	Semilog	lnP	=	12.62	-0.04	Q
	SIG	Loglog	lnP	=	44.90	-7.54	lnQ

Table 7.14. Estimated Demand Functions: CBG Case

The characteristics of the demand functions can be seen more clearly with plotted figures shown in Figure 7.21 though Figure 7.25. Although the level of elasticity differ between individual houses and census block case, the shape of the demand functions estimated are very similar for the non-linear cases of fecal for SIG data.

As for secchi depth readings, very inelastic tendency can be seen in very steep sloped demand functions as shown in Figure 7.24 and Figure 7.25. The slope gets steeper dramatically around 240 for the [COR, Semi log] case and around 230 for the [COR, Log log] case. As for SIG data case, semi log and log log functions look very similar to each other and the turning point is around 245.

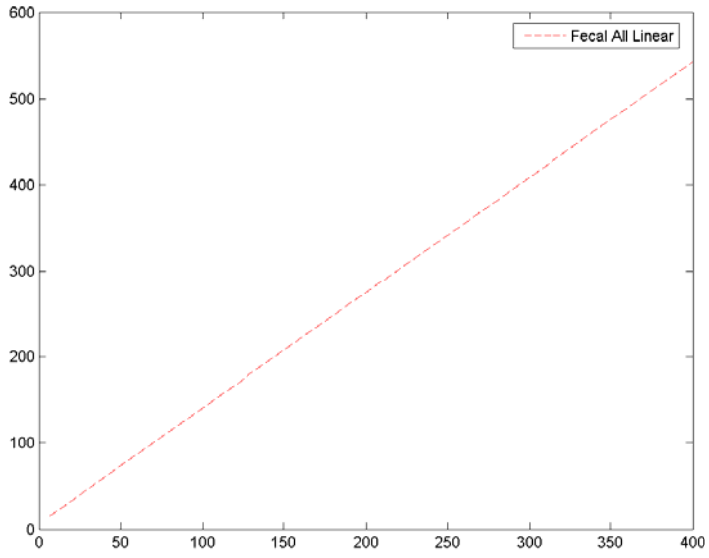


Figure 7.21. Linear Demand Function for Fecal, All Data: CBG Case

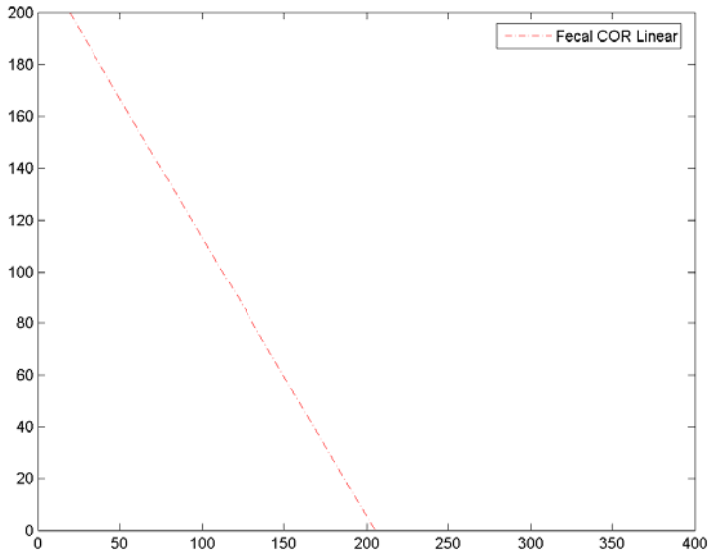


Figure 7.22. Linear Demand Function for Fecal, COR Data: CBG Case

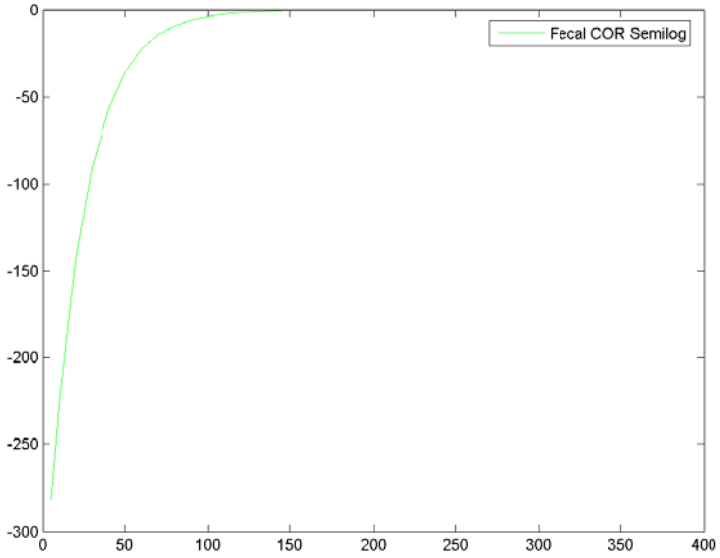


Figure 7.23. Semi Log Demand Functions for Fecal, COR Data: CBG Case

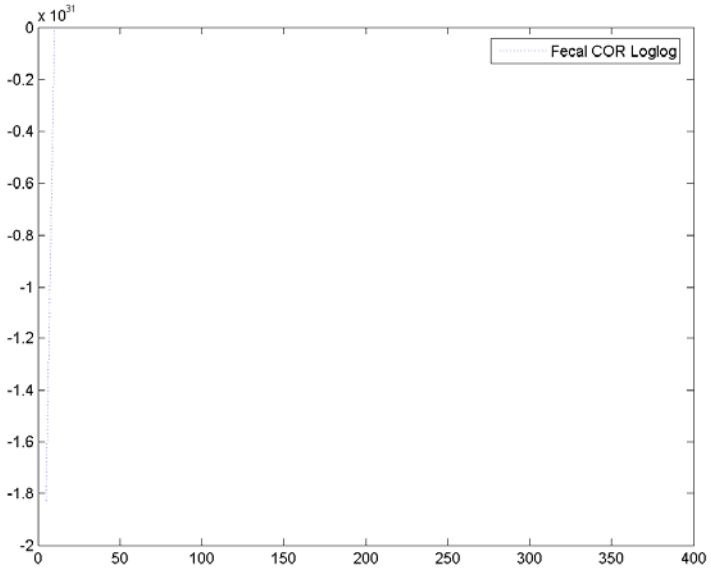


Figure 7.24. Log Log Demand Functions for Fecal, COR Data: CBG Case

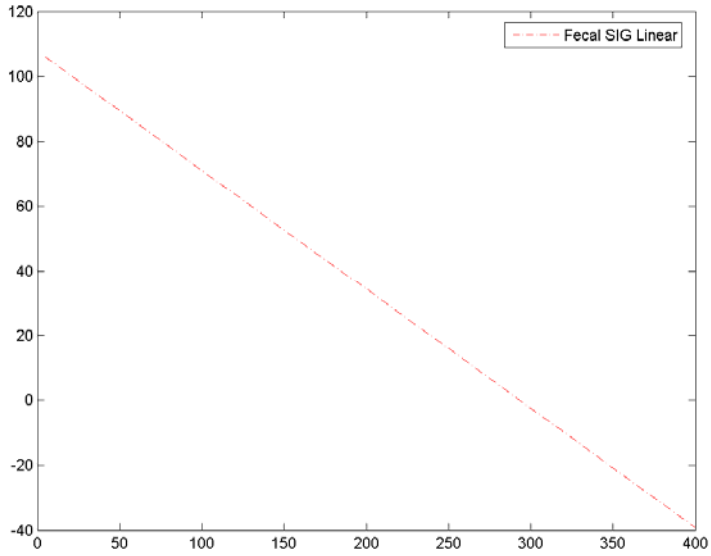


Figure 7.25. Linear Demand Functions for Fecal, SIG Data: CBG Case

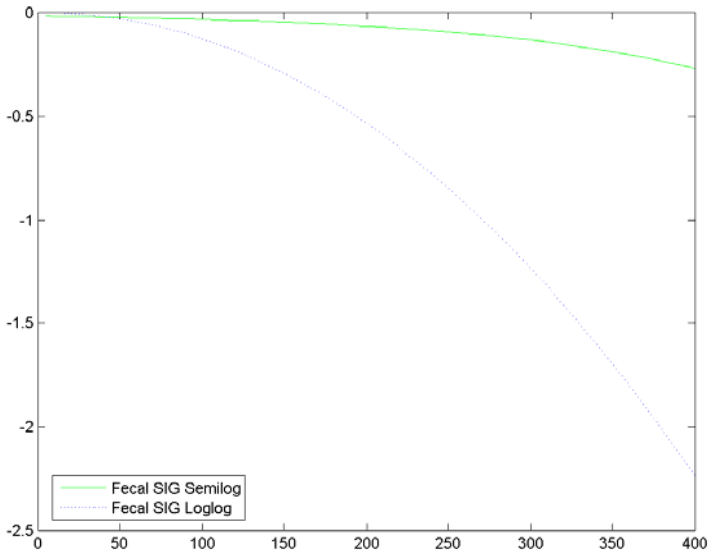


Figure 7.26. Non-linear Demand Functions for Fecal, SIG Data: CBG Case

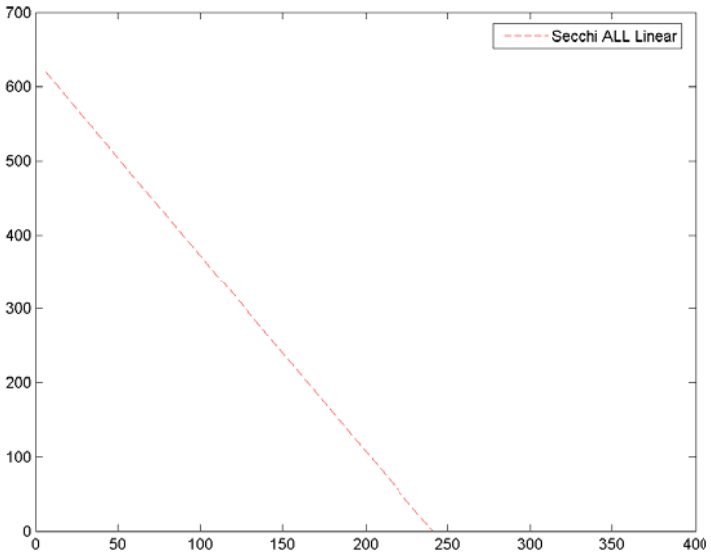


Figure 7.27. Linear Demand Functions for Secchi, All Data: CBG Case

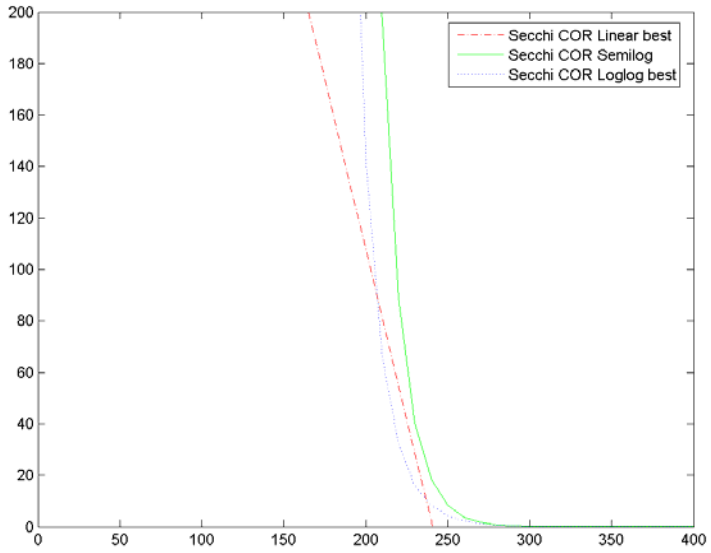


Figure 7.28. Demand Functions for Secchi, COR Data: CBG Case

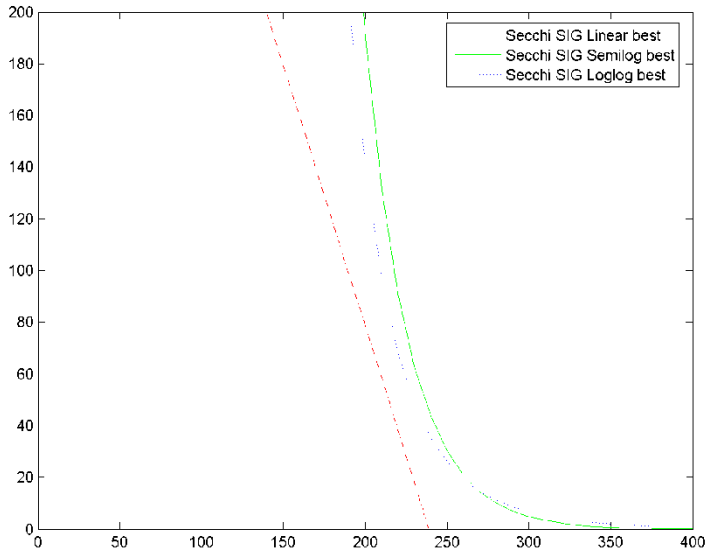


Figure 7.29. Demand Functions for Secchi, SIG Data: CBG Case

7.4.3 Computed Welfare Change: CBG Case

Given the driven inverse demand functions, we calculated non-marginal welfare change by integrating between x axis and the demand curves. The welfare changes for the changes in the level of fecal are very small comparing the case for individual houses. For the case of the [SIG, Log log], it ranges between 2.6 (50 counts change) and 12 (150 counts change) dollars. As before, the welfare changes due to degradation is larger than the case with the change in the opposite direction by the same amount.

				New Fecal Level (From 255 counts /100 ml)							
				Improvement				Degradation			
			N	230	205	155	105	280	305	355	405
	All	Linear	10655	8295	15751	28151	37200	9132	19103	41557	67362
	COR	Linear	7311	-1008	-1343	6	4045	-1680	-4034	-10758	-20173
	SIG	Linear	2275	1231	2733	6548	11445	961	1651	2219	1705
Fecal	COR	Semilog	7311	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SIG	Semilog	2275	-0.9	-1.7	-3.0	-4.1	-1.0	-2.1	-4.8	-8.2
	COR	Loglog	7311	-	-	-	-	-	-	-	-
	SIG	Loglog	2275	-2.6	-5.0	-9.0	-12.0	-2.9	-6.0	-12.9	-20.7

Table 7.15 Computed Welfare Change for Fecal (in $ 1996) : CBG Case

				New Secchi Level (From 220 cm)							
				Improvement				Degradation			
			N	245	270	320	370	195	170	120	70
	All	Linear	10655	546	-559	-7722	-21488	2834	6044	18690	37940
	COR	Linear	10655	546	-559	-7722	-21488	2834	6044	18690	37940
	SIG	Linear	7193	328	-602	-6237	-16906	2055	4432	13896	28395
Secchi	COR	Semilog	10655	967	1100	1120	1121	10988	58054	3.E+06	2.E+08
	SIG	Semilog	7193	1484	2075	2403	2455	4985	13110	95959	619520
	COR	Loglog	10655	381	457	479	481	3604	20213	3.E+06	9.E+09
	SIG	Loglog	7193	1159	1693	2096	2217	3691	10096	118690	4.E+06

Table 7.16 Computed Welfare Change for Secchi (in $ 1996) : CBG Case

The calculated results can be found in Table 7.16 for the water clarity. The welfare changes computed is almost 20 times larger than the case with individual houses. This is mainly due to the very inelastic demand function estimated. We again found that the welfare changes for COR data is smaller than the case with SIG data. The degradation of water clarity changes households' welfare with larger magnitude than the case of improvement in water clarity.

7.5 Welfare Measure Calculation II

The existing hedonic studies which conducted the second stage hedonic analysis reported a single demand function for a certain good of interest for the entire housing market by controlling socio-demographic features as we just reported in the previous section. As we have shown, the welfare measures we computed are based on the demand function derived by using the mean value of each variable included in the estimation of the demand function. However, it will be a more accurate representation of the welfare changes if we derive the individual demand function for each observation by using individual variables (e.g. MIP for FECAL for house i) and compute the welfare changes for each individual by using the individual demand functions. This means that we do not aggregate the results at any stage of welfare change calculations. For example, as for the ALL data case, we compute 10655 demand functions and integrate over the range between the initial values and the targeted level for all 10655 demand functions separately. In the end, we report the mean value of the calculated individual welfare changes together with other descriptive statistics.

The welfare measures are calculated in the similar settings as the ones reported in the previous sections. However, the initial values of water quality are set to be equal to the actual values individual houses are facing. Therefore, even though we use the same targeted levels of water quality, there may be improvements for some houses and degradations for others since the initial values are different. This fact causes the computed areas under the demand curves to be both negative and positive depending on the relative quantities between the initial water quality and their targeted values. If the targeted value is greater than the initial value, we obtain a positive valued area. On the other hand, if the targeted value is less than the initial value, the computed area will be negative. Therefore, for the case of fecal coliform counts, if the change is an improvement to a household (initial > target), the result will be positive because the inverse demand function for fecal is located in the fourth quadrant, and the value will be negative if the targeted value is greater than the initial value. As for secchi readings, if the change is an improvement with respect to the initial setting (initial < target), we will obtain a positive value as the welfare changes.

After obtaining the welfare changes for each household for each case (targeted values of 105, 155, 205, 230, 280, 305, 355 and 405 for fecal, 70, 120, 170, 195, 245, 270, 320 and 370 for secchi), we first compute the overall mean for each case for each data set with three different functional forms as before. This mean is the aggregated mean of the welfare changes for both improvement and degradation of water quality. Therefore, the signs of the means depend on the number of houses that will experience improvements versus the number of houses which will face degradation of water quality. We should

note that this way of identifying the degradation versus improvements does not work for the linear specifications of the demand functions since, especially with the inelastic demand, the demand line extend to the fourth quadrant from the first, causing calculated area to include the areas both above and below the x axis. Therefore, for the case of linear demand functions, signs of the welfare changes computed do not necessarily indicate whether the house is experiencing an improvement or a degradation. Although we report the computed values, these values should be seen just as a reference.

In order to further observe the composition of the positive and negative influences, we computed means only with positive areas and with negative areas separately, together with some descriptive statistics such as standard deviation, minimum, maximum and counts. The derivations of the integral under each demand curve are done by Matlab, and the welfare changes calculated are reported in Tables from 7.17 to 7.30 for both fecal coliform and secchi cases. In the tables, the second row states the type of the data set and the functional form of the demand function used, and the third row indicates the targeted level of water quality, the counts per 100 ml for fecal and the centimeters for secchi. The fourth row shows the overall mean of welfare changes, and the calculated means for negative and positive welfare changes are reported in the fifth and the sixth rows, respectively. The rest of the table includes standard deviation, minimum, maximum and the counts for all three cases (ALL,NEG, POS).

		FECAL ALL Linear							
		105	155	205	230	280	305	355	405
MEAN	ALL	24681	26273	26507	26650	27236	27830	30378	35166
	NEG	-10787	-11375	-13635	-14813	-16876	-17654	-17934	-16800
	POS	194793	159774	132828	118823	94658	83934	68823	60411
STD.DEV.	ALL	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05
	NEG	8365	8786	11397	12500	13550	13601	13399	12905
	POS	5.E+05	5.E+05	4.E+05	4.E+05	4.E+05	4.E+05	3.E+05	3.E+05
MIN	ALL	-1.E+05	-8.E+04	-6.E+04	-6.E+04	-7.E+04	-8.E+04	-8.E+04	-8.E+04
	NEG	-1.E+05	-8.E+04	-6.E+04	-6.E+04	-7.E+04	-8.E+04	-8.E+04	-8.E+04
	POS	5	0	8	0	0	1	11	1
MAX	ALL	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06
	NEG	-3	-2	-5	-1	-1	-1	-3	-11
	POS	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06	4.E+06
COUNT	ALL	10665	10665	10665	10665	10665	10665	10665	10665
	NEG	8825	8319	7742	7356	6447	5890	4726	3487
	POS	1840	2346	2923	3309	4218	4775	5939	7178

Table 7.17. Mean Welfare Changes: Individual Demand Functions, Fecal, ALL, Linear

		FECAL COR Linear							
		105	155	205	230	280	305	355	405
MEAN	ALL	-7801	-9673	-10839	-11309	-12221	-12623	-13955	-16295
	NEG	-126865	-97210	-74395	-69254	-54905	-49302	-38516	-33208
	POS	9500	8525	8500	8492	8816	9104	9862	10622
STD.DEV.	ALL	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05
	NEG	4.E+05	4.E+05	3.E+05	3.E+05	3.E+05	2.E+05	2.E+05	2.E+05
	POS	1.E+04	8.E+03	7.E+03	8.E+03	9.E+03	9.E+03	1.E+04	1.E+04
MIN	ALL	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06
	NEG	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06	-2.E+06
	POS	1	0	17	2	0	2	0	3
MAX	ALL	1.E+05	1.E+05	1.E+05	9.E+04	7.E+04	7.E+04	9.E+04	9.E+04
	NEG	-13	-1	-1	-6	-1	-8	-2	-15
	POS	1.E+05	1.E+05	1.E+05	9.E+04	7.E+04	7.E+04	9.E+04	9.E+04
COUNT	ALL	5328	5328	5328	5328	5328	5328	5328	5328
	NEG	676	917	1243	1357	1759	1982	2623	3272
	POS	4652	4411	4085	3971	3569	3346	2705	2056

Table 7.18 Mean Welfare Changes: Individual Demand Functions: Fecal, COR, Linear

		FECAL COR SemiLog							
		105	155	205	230	280	305	355	405
MEAN	ALL	-11572	-43479	-46739	-46991	-47097	-47106	-47109	-47110
	NEG	-115626	-92417	-93239	-86146	-82526	-69719	-70288	-58865
	POS	34699	3975	394	131	12	5	0	0
STD.DEV.	ALL	4.E+05	5.E+05	5.E+05	5.E+05	5.E+05	5.E+05	5.E+05	5.E+05
	NEG	8.E+05	7.E+05	7.E+05	7.E+05	6.E+05	6.E+05	6.E+05	5.E+05
	POS	93083	9378	960	317	32	11	1	0
MIN	ALL	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07
	NEG	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07	-2.E+07
	POS	0	0	0	0	0	0	0	0
MAX	ALL	2.E+06	2.E+05	2.E+04	5303	508	146	15	2
	NEG	-1	-1	-1	-1	-1	0	-1	0
	POS	2.E+06	2.E+05	2.E+04	5303	508	146	15	2
COUNT	ALL	5328	5328	5328	5328	5328	5328	5328	5328
	NEG	1640	2623	2682	2910	3041	3600	3571	4264
	POS	3688	2705	2646	2418	2287	1728	1757	1064

Table 7.19 Mean Welfare Changes: Individual Demand Functions: Fecal, COR, Semilog

		FECAL COR Loglog							
		105	155	205	230	280	305	355	405
MEAN	ALL	-2.E+24	-2.E+24	-2.E+24	-2.E+24	-2.E+24	-2.E+24	-2.E+24	-2.E+24
	NEG	-8.E+24	-5.E+24	-5.E+24	-5.E+24	-4.E+24	-3.E+24	-4.E+24	-3.E+24
	POS	1.E+17	4.E+12	2.E+09	9.E+07	2.E+05	3.E+04	3.E+02	2.E+01
STD.DEV.	ALL	1.E+26	1.E+26	1.E+26	1.E+26	1.E+26	1.E+26	1.E+26	1.E+26
	NEG	3.E+26	2.E+26	2.E+26	2.E+26	2.E+26	2.E+26	2.E+26	2.E+26
	POS	5.E+18	1.E+14	6.E+10	3.E+09	9.E+06	1.E+06	1.E+04	5.E+02
MIN	ALL	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28
	NEG	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28	-1.E+28
	POS	-5.E-01	-5.E-01	-5.E-01	-5.E-01	-5.E-01	9.E-25	-5.E-01	4.E-28
MAX	ALL	2.E+20	5.E+15	3.E+12	1.E+11	4.E+08	4.E+07	7.E+05	2.E+04
	NEG	-6.E-01	-5.E-01	-5.E-01	-5.E-01	-5.E-01	-7.E-37	-5.E-01	-7.E-37
	POS	2.E+20	5.E+15	3.E+12	1.E+11	4.E+08	4.E+07	7.E+05	2.E+04
COUNT	ALL	5328	5328	5328	5328	5328	5328	5328	5328
	NEG	1594	2477	2484	2671	2734	3600	2823	4264
	POS	3734	2851	2844	2657	2594	1728	2505	1064

Table 7.20 Mean Welfare Changes: Individual Demand Functions: Fecal, COR, Loglog

		FECAL SIG Linear							
		105	155	205	230	280	305	355	405
MEAN	ALL	-1268	-3055	-4503	-5019	-5813	-5969	-6438	-7097
	NEG	-146892	-147832	-90669	-71335	-53063	-48305	-40084	-32807
	POS	9966	8322	7448	7328	7272	5809	7934	8532
STD.DEV.	ALL	9.E+04	9.E+04	9.E+04	9.E+04	1.E+05	1.E+05	1.E+05	1.E+05
	NEG	3.E+05	3.E+05	3.E+05	2.E+05	2.E+05	3.E+05	2.E+05	2.E+05
	POS	1.E+04	8.E+03	7.E+03	7.E+03	8.E+03	8.E+03	9.E+03	1.E+04
MIN	ALL	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06
	NEG	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06	-1.E+06
	POS	548	33	11	5	0	-5701	0	2
MAX	ALL	2.E+05	1.E+05	1.E+05	1.E+05	1.E+05	9.E+04	8.E+04	8.E+04
	NEG	523	-44	-12	-11	-1	-6185	-1	-5
	POS	2.E+05	1.E+05	1.E+05	1.E+05	1.E+05	9.E+04	8.E+04	8.E+04
COUNT	ALL	3555	3555	3555	3555	3555	3555	3555	3555
	NEG	254	259	433	558	771	254	1064	1344
	POS	3301	3296	3122	2997	2784	3301	2491	2211

Table 7.21 Mean Welfare Changes: Individual Demand Functions: Fecal, SIG, Linear

		FECAL SIG SemiLog							
		105	155	205	230	280	305	355	405
MEAN	ALL	2159689	2159687	2159685	2159684	2159680	2159678	2159672	2159663
	NEG	-1	-3	-6	-8	-11	-11	-18	-26
	POS	3079701	3572682	3687651	3860074	4556489	6231884	6861195	9086002
STD.DEV.	ALL	2.E+07	2.E+07	2.E+07	2.E+07	2.E+07	2.E+07	2.E+07	2.E+07
	NEG	2	3	6	8	12	14	21	29
	POS	3.E+07	3.E+07	3.E+07	3.E+07	3.E+07	4.E+07	4.E+07	5.E+07
MIN	ALL	-29	-66	-118	-151	-238	-295	-442	-651
	NEG	-29	-66	-118	-151	-238	-295	-442	-651
	POS	0	0	0	0	0	0	0	0
MAX	ALL	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08
	NEG	-1	-1	-1	-1	-1	0	-1	0
	POS	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08	3.E+08
COUNT	ALL	3555	3555	3555	3555	3555	3555	3555	3555
	NEG	1062	1406	1473	1566	1870	2323	2436	2710
	POS	2493	2149	2082	1989	1685	1232	1119	845

Table 7.22 Mean Welfare Changes: Individual Demand Functions: Fecal, SIG, Semilog

We are going to discuss the results for the SIG data set with the log log demand function cases as an example. The number of houses experiencing improvements (degradation) increases as the target level of fecal counts becomes lower (higher). Overall means indicate that the welfare changes for targeting the new levels of fecal counts increase the welfares of households by 341 dollars if the target is 105 counts and it is 252 dollars if it is 405 counts. By targeting 205 counts, for example, 2088 houses are affected positively (welfare increase) and 1467 houses are affected negatively (welfare decrease), meaning the initial fecal levels were lower than 205 for these 1467 houses. The average welfare changes for the houses facing the degradation vary from 3 dollars (for 105 counts) to 99 dollars (for 405 counts). On the other hand, for the houses which are going to experience the improvements of fecal counts to the targeted level, the welfare improvements range between 481 and 1379 dollars.

If we can assume that the population of the SIG data set is computed by the percentage of SIG to ALL data (33%) times real population of four counties (466,992 in 2000), the welfare increase for targeting, for example 155 counts can be computed as 0.33*466992*564, or 86,916,551 dollars (all amounts are in 1996 dollars) and the welfare decreases for the same level are computed as 0.33*466992*8, or 1,232,859 dollars. If we use the overall average value of 337 dollars, the net welfare gain is computed as 51,934,180 dollars.

		FECAL SIG Loglog							
		105	155	205	230	280	305	355	405
MEAN	ALL	341	337	331	326	313	304	282	252
	NEG	-3	-8	-19	-25	-41	-45	-69	-99
	POS	481	564	577	615	707	963	1125	1379
STD.DEV.	ALL	3193	3193	3193	3193	3194	3195	3197	3203
	NEG	6	20	48	66	113	134	206	298
	POS	3782	4093	4149	4288	4608	5363	5804	6422
MIN	ALL	-166	-647	-1585	-2275	-4199	-5473	-8750	-13137
	NEG	-166	-647	-1585	-2275	-4199	-5473	-8750	-13137
	POS	0	0	0	0	0	0	0	0
MAX	ALL	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04
	NEG	-1	-1	-1	-1	-1	0	-1	0
	POS	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04	5.E+04
COUNT	ALL	3555	3555	3555	3555	3555	3555	3555	3555
	NEG	1033	1408	1467	1603	1871	2323	2510	2710
	POS	2522	2147	2088	1952	1684	1232	1045	845

Table 7.23 Mean Welfare Changes: Individual Demand Functions: Fecal, SIG, Loglog

		SECCHI ALL Linear							
		70	120	170	195	245	270	320	370
MEAN	ALL	31907	11456	913	-152	-524	-1494	-9253	-24068
	NEG	-13066	-22679	-24066	-15005	-8613	-7032	-10875	-24400
	POS	33746	14265	4380	4444	-1794	6231	6461	6361
STD.DEV.	ALL	1.E+04	1.E+04	1.E+04	1.E+04	1.E+04	1.E+04	9380	8026
	NEG	9.E+03	1.E+04	2.E+04	2.E+04	1.E+04	1.E+04	8062	7375
	POS	9196	6031	3914	4690	9751	5635	6197	6939
MIN	ALL	-3.E+04	-5.E+04	-6.E+04	-7.E+04	-7.E+04	-6.E+04	-5.E+04	-4.E+04
	NEG	-3.E+04	-5.E+04	-6.E+04	-7.E+04	-7.E+04	-6.E+04	-5.E+04	-4.E+04
	POS	142	40	10	3	-26537	1	0	25
MAX	ALL	1.E+05	7.E+04	4.E+04	4.E+04	4.E+04	4.E+04	4.E+04	3.E+04
	NEG	-15	-15	-1	-1	2717	0	-8	-44
	POS	1.E+05	7.E+04	4.E+04	4.E+04	4.E+04	4.E+04	4.E+04	3.E+04
COUNT	ALL	10665	10665	10665	10665	10665	10665	10665	10665
	NEG	419	811	1300	2520	4987	6212	9667	10550
	POS	10246	9854	9365	8145	5678	4453	998	115

Table 7.24 Mean Welfare Changes: Individual Demand Functions: Secchi, ALL, Linear

		SECCHI COR Linear							
		70	120	170	195	245	270	320	370
MEAN	ALL	19127	8237	2278	1549	1393	1219	-1476	-7440
	NEG	-3701	-7306	-9409	-6894	-4242	-3177	-4118	-8984
	POS	19439	9029	3366	3044	3953	4380	5320	6960
STD.DEV.	ALL	7595	6150	5077	5373	6376	6393	6015	6478
	NEG	2852	5007	6914	7518	6769	5987	3553	3953
	POS	7152	5054	3134	3019	4163	4540	5716	7611
MIN	ALL	-1.E+04	-2.E+04	-3.E+04	-3.E+04	-3.E+04	-3.E+04	-2.E+04	-2.E+04
	NEG	-1.E+04	-2.E+04	-3.E+04	-3.E+04	-3.E+04	-3.E+04	-2.E+04	-2.E+04
	POS	73	28	6	3	0	0	-4	-11
MAX	ALL	73526	51449	32970	28997	37672	40930	53286	62043
	NEG	-71	-12	-1	-1	-1	0	-4	-11
	POS	73526	51449	32970	28997	37672	40930	53286	62043
COUNT	ALL	8754	8754	8754	8754	8754	8754	8754	8754
	NEG	118	424	746	1317	2735	3662	6304	7906
	POS	8636	8330	8008	7437	6019	5092	2451	849

Table 7.25 Mean Welfare Changes: Individual Demand Functions: Secchi, COR, Linear

		SECCHI COR SemiLog							
		70	120	170	195	245	270	320	370
MEAN	ALL	-1127932	-115271	-2220	7240	11457	11810	11967	11985
	NEG	-1127932	-119650	-13699	-4292	-362	-107	-15	-2
	POS	-	135910	22345	25158	18088	14999	13028	12597
STD.DEV.	ALL	1.E+07	1.E+06	1.E+05	1.E+05	1.E+05	1.E+05	1.E+05	1.E+05
	NEG	1.E+07	1.E+06	1.E+05	4.E+04	3566	1174	60	6
	POS	-	2.E+05	1.E+05	2.E+05	2.E+05	2.E+05	1.E+05	1.E+05
MIN	ALL	-5.E+08	-5.E+07	-5.E+06	-2.E+06	-2.E+05	-5.E+04	-1507	-125
	NEG	-5.E+08	-5.E+07	-5.E+06	-2.E+06	-2.E+05	-5.E+04	-1507	-125
	POS	-	465	47	236	3	0	1	0
MAX	ALL	-1.E+04	2.E+06	3.E+06	6.E+06	8.E+06	8.E+06	8.E+06	8.E+06
	NEG	-1.E+04	-1192	-130	-9	-1	0	-1	0
	POS	-	2.E+06	3.E+06	6.E+06	8.E+06	8.E+06	8.E+06	8.E+06
COUNT	ALL	8754	8754	8754	8754	8754	8754	8754	8754
	NEG	8754	8604	5966	5326	3146	1848	712	425
	POS	0	150	2788	3428	5608	6906	8042	8329

Table 7.26 Mean Welfare Changes: Individual Demand Functions: Secchi,COR, Semilog

		SECCHI COR Loglog							
		70	120	170	195	245	270	320	370
MEAN	ALL	-1.E+07	-143315	7288	13840	16583	16866	17045	17087.25
	NEG	-1.E+07	-154782	-9561	-2902	-304	-123	-40	-10.598
	POS	-	514397	43344	39850	26056	21412	18558	17959.7
STD.DEV.	ALL	1.E+08	2.E+06	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05
	NEG	1.E+08	2.E+06	9.E+04	3.E+04	2975	1267	161	36.44849
	POS	-	8.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05	2.E+05
MIN	ALL	-5.E+09	-7.E+07	-4.E+06	-1.E+06	-1.E+05	-5.E+04	-4.E+03	-7.E+02
	NEG	-5.E+09	-7.E+07	-4.E+06	-1.E+06	-1.E+05	-5.E+04	-4.E+03	-7.E+02
	POS	-	973	42	160	2	0	2	0.119871
MAX	ALL	-2.E+05	7.E+06	8.E+06	8.E+06	8.E+06	8.E+06	8.E+06	8.E+06
	NEG	-2.E+05	-2380	-103	-7	-1	0	-2	-0.76751
	POS	-	7.E+06	8.E+06	8.E+06	8.E+06	8.E+06	8.E+06	8.E+06
COUNT	ALL	8754	8754	8754	8754	8754	8754	8754	8754
	NEG	8754	8604	5966	5326	3146	1848	712	425
	POS	0	150	2788	3428	5608	6906	8042	8329

Table 7.27 Mean Welfare Changes: Individual Demand Functions: Secchi, COR, Loglog

		SECCHI SIG Linear							
		70	120	170	195	245	270	320	370
MEAN	ALL	16901	7971	3195	2556	2554	2662	1352	-2320
	NEG	-2607	-4698	-7151	-6097	-3290	-2361	-2622	-5340
	POS	17003	8337	3714	3245	3906	4476	5455	7119
STD.DEV.	ALL	7784	5977	4217	4178	5256	5585	6141	7110
	NEG	2370	3421	4459	5285	4930	4371	2515	2751
	POS	7673	5625	3469	3193	4321	4811	6081	8206
MIN	ALL	-9651	-2.E+04	-2.E+04	-2.E+04	-2.E+04	-2.E+04	-2.E+04	-1.E+04
	NEG	-9651	-2.E+04	-2.E+04	-2.E+04	-2.E+04	-2.E+04	-2.E+04	-1.E+04
	POS	48	3	18	4	0	0	1	3
MAX	ALL	8.E+04	5.E+04	3.E+04	3.E+04	4.E+04	5.E+04	6.E+04	7.E+04
	NEG	-31	-23	-42	-11	-1	-1	-1	-2
	POS	8.E+04	5.E+04	3.E+04	3.E+04	4.E+04	5.E+04	6.E+04	7.E+04
COUNT	ALL	5796	5796	5796	5796	5796	5796	5796	5796
	NEG	30	163	277	427	1089	1538	2944	4391
	POS	5766	5633	5519	5369	4707	4258	2852	1405

Table 7.28 Mean Welfare Changes: Individual Demand Functions: Secchi, SIG, Linear

		SECCHI SIG SemiLog							
		70	120	170	195	245	270	320	370
MEAN	ALL	-56969	-16314	-2130	981	3905	4546	5148	5358
	NEG	-56969	-17090	-5937	-2769	-630	-363	-219	-58
	POS	-	12913	4667	6426	6186	5599	5532	5628
STD.DEV.	ALL	3.E+05	1.E+05	2.E+04	1.E+04	2.E+04	3.E+04	3.E+04	3.E+04
	NEG	3.E+05	1.E+05	3.E+04	6999	949	531	198	56
	POS	-	1.E+04	1.E+04	2.E+04	3.E+04	3.E+04	3.E+04	3.E+04
MIN	ALL	-2.E+07	-7.E+06	-1.E+06	-2.E+05	-2.E+04	-4701	-1422	-354
	NEG	-2.E+07	-7.E+06	-1.E+06	-2.E+05	-2.E+04	-4701	-1422	-354
	POS	-	116	39	162	14	3	18	3
MAX	ALL	-5431	1.E+05	2.E+05	5.E+05	1.E+06	1.E+06	2.E+06	2.E+06
	NEG	-5431	-418	-197	-19	-4	-2	-59	-17
	POS	-	1.E+05	2.E+05	5.E+05	1.E+06	1.E+06	2.E+06	2.E+06
COUNT	ALL	5796	5796	5796	5796	5796	5796	5796	5796
	NEG	5796	5646	3715	3432	1940	1024	387	275
	POS	0	150	2081	2364	3856	4772	5409	5521

Table 7.29 Mean Welfare Changes: Individual Demand Functions: Secchi, SIG, Semilog

		SECCHI SIG Loglog							
		70	120	170	195	245	270	320	370
MEAN	ALL	-111286	-15404	-1263	1207	3539	4121	4792	5138
	NEG	-111286	-16372	-4785	-2271	-627	-448	-391	-147
	POS	-	21028	5025	6255	5635	5101	5163	5402
STD.DEV.	ALL	7.E+05	1.E+05	2.E+04	1.E+04	2.E+04	2.E+04	2.E+04	3.E+04
	NEG	7.E+05	1.E+05	2.E+04	5522	950	673	358	142
	POS	-	2.E+04	1.E+04	2.E+04	2.E+04	22976	25293	26972
MIN	ALL	-5.E+07	-7.E+06	-1.E+06	-1.E+05	-2.E+04	-5834	-2576	-894
	NEG	-5.E+07	-7.E+06	-1.E+06	-1.E+05	-2.E+04	-5834	-2576	-894
	POS	-	152	32	128	13	3	25	6
MAX	ALL	-1.E+04	1.E+05	2.E+05	4.E+05	1.E+06	1.E+06	2.E+06	2.E+06
	NEG	-1.E+04	-543	-164	-16	-4	-2	-94	-41
	POS	-	1.E+05	2.E+05	4.E+05	1.E+06	1.E+06	2.E+06	2.E+06
COUNT	ALL	5796	5796	5796	5796	5796	5796	5796	5796
	NEG	5796	5646	3715	3432	1940	1024	387	275
	POS	0	150	2081	2364	3856	4772	5409	5521

Table 7.30 Mean Welfare Changes: Individual Demand Functions: Secchi, SIG, Loglog

As for water clarity, if we use the outcomes from the SIG data with the log log specification, all the houses are subject to the welfare loss if the water clarity goes down to the level of 70 centimeters. On the other hand, 95 percent of household included into this data subset experience welfare gains if the water clarity increases to the level of 370 centimeters. If the water clarity target is set to 245 centimeters, 3856 houses are affected positively and 1946 houses face welfare losses. The average welfare gain for the target level of 245 centimeters is 5635 dollars while the average welfare loss is 627 dollars.

If we do the same type of calculation as the fecal case, the percentage of relevant population (54 percent) times the total population (466,992) times 5635 dollars gives total welfare gains for targeting water clarity as 245 centimeters as 1,431,456,925 dollars while the welfare loss is 159,276,574 dollars. The net welfare gain is computed as 899,010,835 dollars.

Once again, we have to emphasize that these computed welfare gains and losses are based on the evaluation of water quality by home purchasers. These monetary values do not include either other services provided by the Lake water quality or benefits/damages to the Lake biology.

The calculated per household average welfare gains and losses could be an indicator for the policy makers to make decisions regarding the water quality controls and welfare changes for various target populations.

7.6 Comparison: Individual Houses vs. Census Block Group Case

The difference between individual houses and census block group case is the unit of building blocks that we used for clustering. From different clustering practice, we produced six relatively similar clusters between two cases and four obviously different clusters. We estimated separate first stage and second stage hedonic models, derived demand functions and calculated welfare measures for the change in water quality variable.

Regardless of the similarities in clustering outcomes, the estimated inverse demand functions for each case are quite different in terms of intercepts and the price elasticity of demand. Due to very inelastic demand functions derived for Secchi case, the calculated welfare changes due to the changes in the same amount of water quality differ quite significantly.

7.7 Conclusion

The second stage hedonic analysis has been conducted and reported in this chapter. The outcomes for fecal coliform demand estimation do not look credible especially for COR data set because of very low adjusted R-squares and non-robustness in signs of estimated coefficients. It is highly likely that the unstable results for the fecal variable are due to one or more omitted variable(s) which we do not have at hand, but influencing the demand for fecal in great deal. This may be correlated with a certain factor related to the costs of reducing the discharge of organic matter from the houses although we cannot prove this possibility at this point.

In most of the cases, we found that fecal coliform and water clarity are substitutes to each other. Although the distance to the beach variable was mostly statistically insignificant, in some cases it was found to be a complement to the water quality.

If we compare the results for secchi in both individual houses and census block group cases, the more consistent results across different specifications are found for individual houses case. The estimated demand functions also look more reasonable for individual houses case with less inelastic results. The comparison between fecal and secchi variables tell us that fecal coliform value itself may not be well reflected to housing price to reveal its value. Considering the more robust results for secchi readings, it has been confirmed that what people observe when they make house purchase decision is not actual bacterial counts level, but water clarity and probably beach closing information which we could not incorporate due to lack of exact data and variability of the data.

For all the cases, we found that the welfare changes from the improvements in water quality are less than the change is the same amount of water quality in the other direction, degradation. We also found that estimated welfare changes are larger for the houses whose housing value is significantly affected by water quality comparing to the houses whose value may be affected, but not statistically significantly.

Computed welfare change for the improvement of water clarity by 50 centimeters (from 220 to 270) is 104 dollars for individual houses case and 1693 dollars for census block group case while it increase to176 dollars and 2217 dollars for the increase by 1.5 meters for individual houses and census block group cases, respectively for the houses whose values are significantly affected by water clarity. If water clarity decrease by 50

centimeters, the welfare lost is estimated as 101 dollars and 3691 dollars for individual houses and census block group cases, respectively. Due to the low elasticity, the degree of welfare changes increase substantially for the changes beyond the steep changes in the slope of the demand functions.

We further analyzed the welfare changes by using demand functions derived specifically for each household. Welfare changes based on the individual demand functions were computed by integrating under each demand curve for multiple scenarios. If we consider our SIG Fecal data represents 33 percent of entire population in four counties, the total estimated net benefit was derived as 51,934,180 dollars for targeting 155 fecal coliform counts. The total net welfare gain was computed as 899,010,835 dollars for targeting 245 centimeters of water clarity.

CHAPTER 8

CONCLUSION AND FUTURE WORKS

Demand functions of water quality of Lake Erie in terms of fecal coliform counts and secchi disk depth readings are estimated by using Cluster Analysis, the first and the second stages of hedonic price models. Two types of Cluster Analysis are implemented by using individual houses and census block group as building block of each cluster. Four similarity measures (CDF transformation, CDF + Hamming, CDF + Categorical 1, CDF + Categorical 2) are employed in order to handle mixed cluster variables type (continuous and categorical) more properly comparing to (standardized) Euclidean distances often used in the determination of market segmentations. Clustering outcomes from four different settings are compared in terms of computed weighted mean squared errors (WMSE) from ordinary least squares estimation for each cluster. The optimal numbers of clusters are determined by identifying the "knee-point" from WMSE plots together with the information obtained from weight R-squares calculation. For individual houses case, Categorical 1 method with 11 clusters is adopted while for census block group case, Categorical 2 method with 10 clusters is chosen for the following hedonic estimations.

As the result of Chow test, two clusters in individual houses case are merged together. Therefore, for both settings, we used 10 clusters as determined submarkets in the extent of our data.

Given determined clusters, we estimated spatial hedonic price models after confirming the type of spatial autoregressive models by using robust Lagrange Multiplier tests with four types of weight matrices. Spatial error model has been tested more likely model for all clusters. We observe mixed signs for the influence of fecal coliform counts on housing price from spatial hedonic price estimations. Therefore, we define two different subsets of data in order to analyze the influence of water quality to houses which are influenced by water quality in different degree. The first subsets of data include the houses which are influenced negatively by fecal coliform counts and positively by water clarity regardless of the significance (COR Data). The second subsets of data include the observations that are affected by fecal coliform negatively and by water clarity positively at statistically significant level (SIG Data).

Average marginal implicit prices estimated from the first stage of estimation are minus 21.6 dollars for fecal in the case with COR data. For SIG data, it is minus 30.5 dollars. The estimated MIP for secchi of SIG data is 40.5 dollars and it is 56 dollars for SIG data. Note that all the prices used and derived in our study is in year 1996 dollar. For the results from census block group case, MIPs for fecal are estimated as minus 18 dollars for COR data and minus 53.6 dollars for SIG data. MIPs for secchi is derived as 33.9 dollars and 43 dollars for COR and SIG data, respectively.

Based on the estimated MIPs for each observation, we estimated the second stage of hedonic estimation by using two-stage least squares. Estimated results suggest that fecal coliform counts is a substitute to water clarity in house owner's demand determination. Given the estimated results from two stage least squares, we computed inverse demand functions and derived welfare measures for non-marginal changes in water quality. The magnitudes of welfare changes are greater for the water quality degradation comparing to the improvements. Welfare changes are larger for the case of SIG data set comparing to COR data set. Estimated welfare change for water clarity by using SIG data and log log specification is 63, 104, 151 and 176 dollars for the improvements of water clarity by 25, 50, 100 and 150 centimeters from 2.2 meters, respectively, while they are 101, 276, 1272 and 8031 dollars for the decreases in water clarity by 25, 50, 100 and 150, respectively. Due to the very inelastic nature of demand functions estimated for secchi in census block group case, the derived welfare measures are about 20 times higher for CBG case.

Although existing studies report aggregated demand function and welfare changes based on the aggregated level of demand function, welfare changes based on individual variables and water quality that each household is facing may represent more accurate welfare changes. Therefore, we derived the demand functions for each observation and computed welfare changed based on the individual demand functions. If we consider our SIG Fecal data represents 33 percent of entire population in four counties, the total estimated net benefit was derived as 51,934,180 dollars for targeting 155 fecal coliform counts. The total net welfare gain was computed as 899,010,835 dollars for targeting 245 centimeters of water clarity.

As a future work, we would like to explore more about optimal clustering in the area of hedonic price analysis. In this study, we used six clustering variables (median household income, distance to the coast line, distance to the closest city, x, y coordinates, and categorical value indicating each municipality which include cites, villages and townships. Although we used equal weights to all these variables, it is possible that different housing submarket have different priority variables which have higher weights in the determination of submarkets. For example, houses near the coast line may have different priority for being closer to a city comparing to the houses near the city center. Figure out these weights have to be done simultaneously with determining the cluster boundaries. Finding the "right" amount of variables that are the good representative of given data and assigning the weights on each attribute depending on their importance are the problem dealt in the area of Feature Selection. Due to the complexity of the process, we place this task as our future work.

Tying and comparing different methods for the determination of the optimal number of clusters is also interesting thing to do. There are various methods to do this task in different academic fields. However, the method has to be adjusted to reflect data characteristics at hand. Therefore studies to determine the method will be necessary in market segmentation area.

We used Kriging in order to handle secchi disk depth readings data. However, more ideal interpolation should include the modeling of water flows from rivers in terms of direction and magnitude, coastal geography, weather conditions and other related biological situations. We leave the incorporation of such modeling into the analysis as future work.

BIBLIOGRAPHY

Aksoy, S., and Haralick, R.M. (2001). "Feature Normalization and Likelihood-Based Similarity Measures for Image Retrieval," *Pattern Recognition Letters*, 22(5):563-582, May.

Anselin, L. and Bera, A.K. (1998), "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics", in Ullah, A. and D.E.A. Giles, eds., *Handbook of Applied Economic Statistics*, New York, Marcel Dekker

Anselin, L, Bera, A.K., Florax, R., Yoon, M.J. (1996),"Simple Diagnostic tests for spatial dependence." *Regional Science and Urban Economics*, 26: 77-104.

Anselin, L. (2004), "Spatial Interpolation and Measures of the Effect of Air Quality in Hedonic House Price Models." presented at CURA roundtable, the Ohio State University on Oct.18.2004.

Arguea, N.M and Hsiao, C., (2000). "Market Values of Environmental Amenities: A Latent Variable Approach." *Journal of Housing Economics*. 9: 104-126.

Baltagi, B.H. and Li, D. (2001), " LM tests for Functional Form and Spatial Error Correlation." *International Regional Science Review*, 24, 2: 194-225.

Baltagi, B.H. and Li, D. (2003), "Testing for Linear and Log-Linear Models Against Box-Cox Alternatives with Spatial Lag Dependence." presented at LSU econometrics conference on spatial and spatio temporal econometrics. http://www.spatialstatistics.info/webposting/baltagi/Lagnew2.pdf

Bates, L.K., (2006). "Does Neighborhood Really Matter? Comparing Historically Defined Neighborhood Boundaries with Housing Submarkets." *Journal of Planning Education and Research*, 26:5-17.

Bartik, T.J. (1987), "The Estimation of Demand Parameters in Hedonic Price Models." *Journal of Political Economy*, Vol 95, no1: 81-88.

Bartik, T.J. (1988), "Measuring the benefits of amenity improvements in hedonic price models." *Land Economics,* **64**: 172-183

Beron, K.J., Hanson, Y., Murdoch, J.C.,and Thayer, M.A. (2003),"Hedonic Price Functions and Spatial Dependence: Implications for the Demand for Urban Air Quality. " In *Advances in Spatial Econometrics: Methodology, Tools and Applications*. Edited by Anselin, L., Florax,R.J.G.M, and Rey, S.J. Springer.

Blomquist, G.C., (1974), "The Effect of Electric Utility Power Plant Location on Area Property Value, *Land Economics*, 50:1, 97-100.

Blomquist, G.C., M.C. Berger and J.P. Hoehn, (1988), "New Estimates of Quality of Life in Urban Areas, *American Economic Review,* 78:1, 89-107.

Bourassa, S.C., Hamelink,F., Hoesli, M., and MacGregor, B.D., (1999). "Defining Housing Submarkets" *Journal of Housing Economics*, 8, 160 – 183.

Bourassa, S.C., Hoesli, M., and Vincent, S.P. (2003). "Do Housing Submarkets Really Matter?" *Journal of Housing Economics*, 12, 12-28.

Boyle, M. A. and K. A. Kiel, (2001), "A Survey of House Price Hedonic Studies of the Impact of Environmental Externalities." *Journal of Real Estate Literature*, 9:2, 117-144.

Boyle, K.J. Poor, P.J., and Taylor, L.O. (1999). " Estimating the Demand for Protecting Freshwater Lakes from Eutrophication." *American Journal of Agricultural Economics*. 81, no.5: 1118-1122.

Carrion-Flores, C. and Irwin, E.G., (2004), "Determinants of Residential Land-Use Conversion and Sprawl at the Rural-Urban Fringe." *American Journal of Agricultural Economics,* 86 (4): 889-904.

Cropper, M. L., Deck,L.B. and McConnell, K.E. (1988), "On the choice of functional form for hedonic price functions." *The Review of Economics and Statistics*, 70(4), 668-675.

Dale, L., J. C. Murdoch, M. A. Thayer and P.A. Waddell, (1999), " Do Property Values Rebound from Environmental Stigmas? Evidence from Dallas." *Land Economics*, 75:2, 311-26.

Day, B., (2003). “Submarket Identification in Property Markets: A Hedonic Housing Price Model for Glasgow.” Technical Report, The Centre for Social and Economic Research on the Global Environment, School of Environmental Sciences, University of East Anglia, Norwich, UK.

David, E. L., (1968), “Lakeshore Property Walues: a Guide to Public Investment in Recreation.” *Water Resources Research*, 4:4, 697-707.

Epp, D. J. and K.S. Al-Ani, (1979), “The Effect of Water Quality on Rural Nonfarm Residential Property Values.” *American Journal of Agricultural Economics*, 61:3, 529-34.

Everitt, B.S., Landau, S., and Leese, M. (2001), “Cluster Analysis, Fourth Edition.” Arnold, London, UK.

Feather, T.D. (1992), “ Valuation of Lake Resources Through Hedonic Pricing.” IWR Report 92-R-8. US Army Corps of Engineers, Water Resource Support Center, Institute for Water Resources. Fort Belvoir, VA.

Flower, P.C. and W. R. Ragas, (1994), “The Effects of Refineries on Neighborhood Property Values.” *Journal of Real Estate Research*, 9:3, 319-38.

Freeman, A. M. III. (1993), “The Measurement of Environmental and Resource Values”, Resources for the Future, Washington, DC.

Gamble, H. B. and R. H. Downing. (1982), “Effects of Nuclear Power Plants on Residential Property Values.” *Journal of Regional Science*, 22:4, 457-78.

Goetzmann, W.N., and Wachter, S.M., (1995), “Clustering Methods for Real Estate Portfolios.” *Real Estate Economics*. V23 3: 271-310.

Goodman, A.C., Thibodeau, T.G., (1998), “Housing Market Segmentation.” *Journal of Housing Economics*. 7: 121-143.

Gower, J.C. (1971). “A General Coefficient of Similarity and Some of its Properties.” *Biometrics*, 27, 857-872.

Graves, P., J. C. Murdoch, M.A. Thayer and D. Waidman, (1988), “The Robustness of Hedonic Price Estimation, Urban Air Quality.” *Land Economics*. 64:3, 220-33.

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

Harrison, D. Jr. and D. L. Rubinfeld. (1978), "Hedonic Housing Prices and the Demand for Clean Air." *Journal of Environmental Economics and Management*, 5, 351-68.

Haining, R. (1990), Spatial Data Analysis in the Social and Environmental Science. Cambridge U.K. Cambridge University Press.

Irwin, E.G., (2002), "The Effects of Open Space on Residential Property Values." *Land Economics*, 78 (4): 465-480.

Kahn, S. and Lang, K. (1988), " Efficient Estimation of Structural Hedonic Systems." *International Economic Review*. Vol.29, no.1: 157-166.

Kelejian, H., and Prucha, I.R. (1998), "A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances." *Journal of Real Estate and Finance Economics*. 17: 99-121.

Ketkar, K., (1992), " Hazardous Water Sites and Property Values in the State of New Jersey." *Applied Economics*, 24:6, 647-59.

Kiel, K.A. and K.T. McClain, (1995), "House Prices During Siting Decision Stages: The Case of an Incinerator from Rumor through Operation." *Journal of Environmental Economics and Management*, 28, 241-55.

Kim, C. W., T. T. Phipps, and L. Anselin, (2003), "Measuring the Benefits of Air Quality Improvement: a Spatial Hedonic Approach." *Journal of Environmental Economics and Management*, 45, 24-39.

Kohlhase, J. E., (1991), "The Impact of Toxic Waste Sites on Housing Values." *Journal of Urban Economics*. 30:1, 1-26.

Leggett. C. G. and N.E. Bockstael, (2000), "Evidence of the Effects of Water Quality on Residential Land Prices." *Journal of Environmental Economics and Management*, 39:2,121-44.

Li, M. M. and H. J. brown, (1980), "Micro Neighborhood Externalities and Hedonic Housing Prices." *Land Economics*, 56:2,125-40.

McClelland, G. H., W. D. Schulze and B. Hurd. (1990), "The Effect of Risk Beliefs on Property Values: a Case Study of a Hazardous Waste Site." *Risk Analysis*, 10:485-97.

Mendelsohn, R., D. Hellerstein, M. Huguenin, R. Unsworth and R. Brazee, (1992), "Measuring Hazardous Waste Damages with Panel Models." *Journal of Environmental Economics and Management*. 22:3, 259-71.

Michael, H. J., K.J. Boyle and R. Bouchard, (1996) ,"Water Quality Affects Property Prices; a Case Study of Selected Maine Lakes" Maine Agricultural and Forest Experiment Station, University of Maine, *Miscellaneous Report 398*.

Michaels. R. G. and V. K. Smith, (1990), "Market Segmentation and Valuing Amenities with Hedonic Models: the Case of Hazardous Waste Sites." *Journal of Urban Economics*, 28:2, 223-42.

Murdoch, J. C. and M. A. Thayer, (1988), "Hedonic Price Estimation of Variable Urban Air Quality." *Journal of Environmental Economics and Management*, 15:2, 143-46.

Nelson, A. C. , J. Genereux and M. Genereux, (1992), "Price Effects of Landfills on House Values." *Land Economics*, 68:4, 359-65.

Nelson, J. P. (1978), "Residential Choice, Hedonic Prices and the Demand for Urban Air Quality." *Journal of Urban Economics*, 5:3, 357-69.

Nelson, J. P. (1981), "Three Mile Island and Residential Property Values Empirical Analysis and Policy Implications." *Land Economics*, 57:3, 363-72.

Nelson, G.C. and Hellerstein, D, (1997), "Do Roads Causes Deforestation? Using Satellite Images in Econometric Analysis of Land Use." American Journal of Agricultural Economics, 79: 80-88.

Ohio EPA, Division of Surface Water. 1996. "Ohio Water Resource Inventory. Available online: http://www.epa.state.oh.us/dsw/document_index/305b.html

Palmquist, R. B., (1982), "Estimating the Demand for Air Quality from Property Value Studies." Unpublished paper, Department of Economics and business, North Carolina State University, Raleigh, NC.

Palmquist, R. B. (1984), "Estimating the Demand for the Characteristics of Housing." *The Review of Economics and Statistics*. 66(3):394-404.

Palmquist, R. B. (1988), "Welfare Measurement for Environmental Improvements Using the Hedonic Model: the Case of Nonparametric Marginal Prices." *J. of Environmental Economics and Management.* 15:297 – 312.

Palmquist, R.B., (1991), "Hedonic Methods." in Measuring the Demand for Environmental Quality. Edited by Braden, J.B. and Kolstad, C.D. North Holland, Amsterdam, 77-120.

Papoulis., A. (1991), Probability, Random Variables, and Stochastic Processes. McGraw-Hill, NY, 3rd edition.

Reichert, A. K., M. Small and S. Mohanty, (1992), "The Impact of Landfills on Residential Property Values." *Journal of Real Estate Research*, 7:3, 297-314.

Ridker, R. G. and J. A. Henning, (1967), "The Determinants of Residential Property Values with Special Reference to Air Pollution." *Review of Economics and Statistics*, 49:2, 246-57.

Romesburg, H. C., (1984), Cluster Analysis for Researchers. Lulu Press, North Carolina.

Rosen, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *The Journal of Political Economy*, Vol.82, No.1: 34-55.

Salvador, S., and Chan, P. (2004), "Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms." *16th IEEE ICTAI (International Conference on Tools with Artificial Intelligence)* 15-17 Nov. 2004: 576-584.

Simons, R. A., (1999), "The Effect of Pipeline Ruptures on Noncontaminated Residential Easement Holding Property in Fairfax County." *The Appraisal Journal*, July 255-63.

Steinnes, D. N., (1992), "Measuring the Economics Value of Water Quality." *Annals of Regional Science*, 26, 171-76.

Smith, V. K. and J. C. Huang, (1993), "Hedonic Models and Air Pollution: Twenty-Five Years and Counting." *Environmental Resource Economics*, 3, 381-394.

Smith, V. K. and J. C. Huang (1995), "Can Markets Value Air Quality? A Meta-analysis of Hedonic Property Value Models." *Journal of Political Economics*, 103, 209-227.

Smolen, G. E., G. Moore, and L. V. Conway, (1992), "Economic Effects of Hazardous Chemical and Proposed Radioactive Water Landfills on Surrounding Real Estate Values." *Journal of Real Estate Research* 7:3, 283-95.

Taylor, O. L. (2003), "The Hedonic Method." in A Primer on Nonmarket Valuation, Edited by Champ, A.P., Boyle, J.K., and Brown, C.T. Boston, Kluwer Academic Publishers.

Tiebout, C.M., (1956), "A Pure Theory of Local Expenditures." *Journal of Political Economy,* 64: 416-424.

USEPA (US Environmental Protection Agency), (2006), "Lake Erie Phosphorous." http://www.epa.gov/glnpo/glindicators/water/phosphorusb.html

Watkins, C., (1998), "Property Valuation and the Structure of Urban Housing Market." *Journal of Property Investment and Finance,* Vol. 17 No.2: 157 – 175.

Weand, K. F., (1973), "Air Pollution and Property Values: a Study of the St. Louis Area." *Journal of Regional Science*, 13:1, 91-5.

Young, C. E., (1984), "Perceived Water Quality and the Value of Seasonal Homes." *Water Resources Bulletin*, American Water Association, 20:2 163-68.

Zabel, J. E. and K. A. Kiel, (2000), "Estimating the Demand for Air Quality in Four U.S. Cities." *Land Economics*, 76:2, 174:94.